UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Emely Pujólli da Silva

Facial Expression Recognition in Brazilian Sign Language using Facial Action Coding System

Reconhecimento de expressões faciais na Língua de Sinais Brasileira por meio do sistema de códigos de ação facial

Campinas

2020

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Emely Pujólli da Silva

Facial Expression Recognition in Brazilian Sign Language using Facial Action Coding System

Reconhecimento de expressões faciais na Língua de Sinais Brasileira por meio do sistema de códigos de ação facial

Thesis presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor, in the are of Computer Engineering.

Tese apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Engenharia Elétrica, na Área de Engenharia de Computação.

Superviser (Orientadora): Prof. Dr. Paula Dornhofer Paro Costa

Co-Supervisor (Coorientadora): Prof. Dr. Kate Mamhy Oliveira Kumada

Este exemplar corresponde à versão final da tese defendida pela aluna Emely Pujólli da Silva, e orientada pela Prof. Dr. Paula Dornhofer Paro Costa

_____

Campinas

2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Si38f

Silva, Emely Pujólli da, 1990-
Facial expression recognition in brazilian sign language using facial action coding system / Emely Pujólli da Silva. – Campinas, SP : [s.n.], 2020.

Orientador: Paula Dornhofer Paro Costa.
Coorientador: Kate Mamhy Oliveira Kumada.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Lingua de sinais. 2. Lingua brasileira de sinais. 3. Redes neurais profunda. 4. Visão por computador. I. Costa, Paula Dornhofer Paro, 1978-. II. Kumada, Kate Mamhy Oliveira, 1985-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Reconhecimento de expressões faciais na língua de sinais brasileira por meio do sistema de códigos de ação facial
Palavras-chave em inglês:
Sign language
Brazilian sign language
Deep neural networks
Computer vision
Área de concentração: Engenharia de Computação
Titulação: Doutora em Engenharia Elétrica
Banca examinadora:
Paula Dornhofer Paro Costa [Orientador]
Sandra Eliza Fontes de Avila
Ivani Rodrigues Silva
Sarajane Marques Peres
Neiva de Aquino Albres
Data de defesa: 21-10-2020
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-7745-6151
- Currículo Lattes do autor: http://lattes.cnpq.br/0875535364516080

COMISSÃO JULGADORA – TESE DE DOUTORADO

Candidato: Emely Pujólli da Silva
RA: 151545
Data da Defesa: 21 de outubro de 2020
Título da Tese: “Facial expression recognition in brazilian sign language using facial action coding system”.

Prof. Dra. Paula Dornhofer Paro Costa (FEEC / Unicamp)
Prof. Dra. Sandra Eliza Fontes de Avila (IC / Unicamp)
Prof. Dra. Ivani Rodrigues Silva (FCM / Unicamp)
Prof. Dra. Sarajane Marques Peres (EACH / USP)
Prof. Dra. Neiva de Aquino Albres (CCE / UFSC)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

To my family and to all those who inspired this work.

Acknowledgements

The completion of this thesis could not be possible without the participation and assistance of so many people whose names may not be enumerated. Their contributions are sincerely appreciated and gratefully acknowledge. However, I would like to express my deep appreciation and indebtedness, particularly to the following:

Above all to the Almighty God, the Author of knowledge and wisdom, for his countless love.

To my family, Bernardino, Fatima, and Jolel, who in many ways shared their support, either morally, financially, and physically, thank you. Whenever I've needed them, they were there for everything; I appreciate and love them with all my heart.

Professor Paula Dornhofer Paro Costa, for the affection, support, patience, and, without her guidance and discussions of ideas, this work never would have been done, I am sure. She allowed me to be part of her life, sharing her knowledge with me, and granted me the opportunity to meet amazing people. Paula opened up the door in Engineering for me and, with that, a whole new set of paths.

Professor Kate Mamhy Oliveira Kumada, who, with her ethical gaze as a researcher in the field of applied linguistics, providing me with references and inciting discussions that helped me to mature my linguistic arguments about sign languages.

Deaf community, the creators, and evolutionaries of vibrant sign languages, which without it, this work would not exist. In special, to the Brazilian Deaf community of Brazilian sign language speakers and scholars that show the magnificence and intricate composition of the language.

To the participants of our video corpus, without whom it would be impossible to carry out this research;

To professors, José Mario De Martino and Ivani Rodrigues Silva, coordinators of the Assistive Technologies for the Deaf (TAS) project, allowed my work to participate and thrive within the TAS project atmosphere.

Professor Cícero Lopes Frota and Professor Marcos Valle, for their acceptance in taking me on a scientific math journey. As my previous advisers, they were essential in my training, where I could discover and evolve my passion for research.

To my friends and others who somehow participated in studies, coffee breaks, and group research. This work was not have been the same without them cheering for our

success.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and we also would like to thank them.

I am thrilled to share with the world such a motivating work.

Thank you all.

"A felicidade é viver o presente."
(Eclesiastes)

Abstract

Deaf people around the world use sign languages to communicate but, despite the wide dissemination of such languages, deaf or hard of hearing individuals still face difficulties in communicating with hearing individuals in the absence of an interpreter. Such difficulties negatively impact the access of deaf individuals to an education, the job market, and public services in general. Assistive technology, such as Automatic Sign Language Recognition (ASLR), aims at overcoming such communication obstacles. However, the development of reliable ASLR systems imposes numerous challenges due to the linguistic complexity of sign languages. Sign languages (SLs) are visuospatial linguistic systems that, like any other human language, present global and regional linguistic variations and a grammatical system. Also, sign languages do not rely only on manual gestures but also non-manual markers, such as facial expressions. In SL, facial expressions may differentiate lexical items, participate in syntactic construction, and contribute to change the intensity of a sentence, among other grammatical and affective functions. Associated with the gesture recognition models, facial expression recognition (FER) is an essential component of ASLR technology. In this work, we propose an automatic facial expression recognition (FER) system for the Brazilian Sign Language (Libras). Derived from a literature survey, we present a language study and a different taxonomy for facial expressions of Libras associated with the Facial Action Coding System (FACS). Also, a dataset of facial expressions in Libras was created. An experimental setting was done for the construction of our framework for a preprocessing stage and recognizer model. The features for the classification of the facial actions resulted from applying a combined region of interest and geometric information are given a theoretical basis and better performance than other tested steps. As for classifiers, SqueezeNet returned better accuracy rates. With this, the proposed model's potential comes from the analysis of 77% of the average accuracy of recognition of Libras' facial expressions. This work contributes to the growth of studies that involve the computational vision and recognition aspects of the structure of sign language facial expressions, and its main focus is the importance of facial action annotation in an automated way.

Keywords: Facial Action Unit Recognition; FACS; Non-Manual Markers; Libras.

Resumo

Surdos ao redor do mundo usam a língua de sinais para se comunicarem, porém, apesar da ampla disseminação dessas línguas, os surdos ou indivíduos com deficiência auditiva ainda enfrentam dificuldades na comunicação com ouvintes, na ausência de um intérprete. Tais dificuldades impactam negativamente o acesso dos surdos à educação, ao mercado de trabalho e aos serviços públicos em geral. Tecnologia assistiva, como o Reconhecimento Automático de Língua de Sinais, do inglês Automatic Sign Language Recognition (ASLR), visam superar esses obstáculos de comunicação. No entanto, o desenvolvimento de sistemas ASLR confiáveis apresenta vários desafios devido à complexidade linguística das línguas de sinais. As línguas de sinais (LSs) são sistemas linguísticos visuoespaciais que, como qualquer outra língua humana, apresentam variações linguísticas globais e regionais, além de um sistema gramatical. Além disso, as línguas de sinais não se baseiam apenas em gestos manuais, mas também em marcadores não-manuais, como expressões faciais. Nas línguas de sinais, as expressões faciais podem diferenciar itens lexicais, participar da construção sintática e contribuir para processos de intensificação, entre outras funções gramaticais e afetivas. Associado aos modelos de reconhecimento de gestos, o reconhecimento da expressões faciais é um componente essencial da tecnologia ASLR. Neste trabalho, propomos um sistema automático de reconhecimento de expressões faciais para Libras, a língua brasileira de sinais. A partir de uma pesquisa bibliográfica, apresentamos um estudo da linguagem e uma taxonomia diferente para expressões faciais de Libras associadas ao sistema de codificação de ações faciais. Além disso, um conjunto de dados de expressões faciais em Libras foi criado. Com base em experimentos, a decisão sobre a construção do nosso sistema foi através de pré-processamento e modelos de reconhecimento. Os recursos obtidos para a classificação das ações faciais são resultado da aplicação combinada de uma região de interesse, e informações geométricas da face dado embasamento teórico e a obtenção de desempenho melhor do que outras etapas testadas. Quanto aos classificadores, o SqueezeNet apresentou melhores taxas de precisão. Com isso, o potencial do modelo proposto vem da análise de 77% da acurácia média de reconhecimento das expressões faciais de Libras. Este trabalho contribui para o crescimento dos estudos que envolvem a visão computacional e os aspectos de reconhecimento da estrutura das expressões faciais da língua de sinais, e tem como foco principal a importância da anotação da ação facial de forma automatizada.

Palavras-chaves: Reconhecimento de Unidade de Ação Facial; FACS; Marcadores não-manuais; Libras.

List of Figures

List of Tables

Glossary and List of symbols

A	Assertive
AAM	Active Appearance Model
AD	Action descriptors
ADAM	Adaptative Moment Estimation
AFE	Affective Facial Expression
ANN	Artificial Neural Network
ASLR	Automatic Sign Language Recognition
AU	Action Unit
BSL	British Sign Lanaguage
CC	Condicional Clause
CEP	Research Ethics Committee
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CSL	Chinese Sign Language
d_2	Euclidean distance
DGS	German Sign Language
DQ	Doubt question
DRML	Deep Region Multilabel Learning
EAC	Enhancing and Cropping
ELAN	Eudico Linguistic Annotator
F	Focus
FACS	Facial Action Coding System
FE	Facial Expression
FER	Facial Expression Recognition
FPS	Frames per Second
FR	Facial Recognition
GARN	Generative Adversarial Recognition Network
GEH	Grammatical Facial Expression of Homonymy
GEI	Grammatical Facial Expression of Intensity
GEN	Grammatical Facial Expression of Norm
GES	Grammatical Facial Expression of Sentence
GFE	Grammatical Facial Expression

HMM	Hidden Markov model
HOG	Histogram of Gradients
Libras	Brazilian Sign Language
LSTM	Long Short-Term Memory
MM	Manual Markers
MS	Manual Signs
N	Negative
NME	Non-Manual Expression
NMM	Non-Manual Markers
NN	Neural Network
PAD	Pleasure-Arousal Dominance
PCA	Principal component analysis
RC	Relative Clause
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descendent
SL	Sign Language
T	Topic
TAS	Brazilian Portuguese acronym for *Tecnologias Assistivas para Surdos*
TTS	Text-to-Speech
TTSL	Text-to-Sign Language
UFABC	Federal University of ABC
Unicamp	University of Campinas
WH	WH-Question
WSC	Weakly Supervised Clustering
YN	Yes No Question

Contents

1 Introduction

Facial expression (FE) can be defined as the motion or the change in position of the muscles beneath the face's skin. Through facial expressions, human beings can transmit information. Emotions, feelings, physiological aspects, identity, and language are some visual signs obtained from facial expressions that are a form of nonverbal communication. Facial expression identification and analysis have received special attention from research groups in areas such as linguistics (DACHKOVSKY; SANDLER, 2009; VOS *et al.*, 2009), psychology (DU *et al.*, 2014), pattern recognition (MARTINEZ; VALSTAR, 2016), computer vision (AGIANPUYE; MINOI, 2013), and some areas involved with accessibility (KACORRI, 2015; CARDOSO, 2018). The human face has been the object of study to understand the physiological and psychological aspects of people's behavior during their interaction with their surroundings because part of this interaction is based on facial expressions (EKMAN, 1993; CARDOSO, 2018).

By understanding facial expressions, it is possible to improve the research and development of many applications, like security control, FE synthesis in designing avatars, and even robot-coupled face recognition systems capable of perceiving the emotional states of their operators. Through facial expression analysis, it is possible to recognize emotions, and for instance, to endow software agents with the ability to use this information to improve human-computer interaction (CARDOSO, 2018).

For psychology, humans can adopt a facial expression voluntarily or involuntarily, and there substantial evidence that the face is the primary signal system for showing emotion. Most studies are done around a subset of emotions, generated from the possible emotional relationships and reactions, enough to study and understand relationships displayed by the face. There have been hundreds of experiments on facial expression where the subset of emotions are happiness, surprise, anger, disgust, fear, and sadness. They are called the six basic emotions; often the neutral expression is combined with them and used as an initial reference for analysis of the other expressions (EKMAN; FRIESEN, 2003). Lately, some lines of research are examining pain in terms of facial expression manifestations. Mapping features of pain enables the detection and identification of discomfort, and this information can help prevent suffering or even diagnostic the state of health of an individual (FREITAS-MAGALHÃES, 2013). An interesting way of studying facial expressions is by using a measure to annotate and quickly describe the facial action. The most known model in the literature is called the Facial Action Coding System (FACS) introduced by Ekman e Friesen (1978) in the seventies to code changes

in facial appearance. In this system, the measurements are called Action Units (AUs), which are associated with facial muscles' movements. In addition, facial expressions are used as a kind of non-manual gesture of speech-related information and, in some cases, to replace verbal communication. Studies have shown improvements in using a language when a multimodal approach to analysis is included, i.e., when an investigation related to nonverbal cues is done.

Inclose, facial expressions have a more prominent role in linguistics, where they can compose the channel that transmits information. Structured in gestures are the Sign Languages (SLs), which are visuospatial languages where facial expressions convey grammatical information additionally to the affective state. Even more, their role is fundamental since they withstand the syntactic and semantic structure of the language and also can assume the prosodic expression role. Communication channels in SL include manual signs, multimodal or multi-channel signals, and non-manual markers. Manual signs are composed of hand configuration, movement, hand direction, pivot point, and hands. Multimodal or multi-channel signals are those whose realization requires, besides the use of the hands, actions of other parts of the body (head, face, torso) (XAVIER, 2019). Non-manual markers (NMM) are composed of facial and body expressions. Around the world, deaf people use SL to communicate with others.

According to Stokoe (1960), SLs are grammatically structured and organized into a sentence-forming system, and he positioned facial expressions as constituents elements of the language (STOKOE, 1980). Later on, Baker-Shenk (1985) analyzed the role that facial expression can assume in SL, including their function on giving sense into what is said. When facial expressions transmit the grammatical information in a sentence, they are called grammatical facial expressions (GFE). There are facial expressions that carry emotion or an affective state called affective facial expressions (AFE). The linguistic and affective marks of facial expressions differ in SL in several ways, which marks the different use of the same facial muscles.

The GFE can be found in the morphological levels of the language that assume the role of adjective attribution. Prosody can also take over a position that is marked linguistically by facial expression. For example, imperative can be differentiated from neutral sentences by prosody alone when the command is conveyed by visual prosody on the face. Another prosody usage is to present the same intonational contour of spoken language. At the syntactic level of language, FEs act as one of the structures responsible for building negative, interrogative, affirmative, conditional, relative, topical, and focused. All of these grammatical functions obtained from facial information helps to discriminated and have a significant influence on the automatic recognition implemented to sign language (QUADROS; KARNOPP, 2009; FREITAS *et al.*, 2014; CARDOSO, 2018).

Thus, given the importance of GFE for constructing the discourse in sign language, it becomes clear that applications that aim to process the information transmitted by the speech need to consider the automatic identification of the GFEs. The task of transcribing passages observed within the discourse is called an annotation. Despite being an arduous and time-consuming activity, it is crucial in the advancement of linguistic studies. Efforts by the scientific community are observed to understand and automate characteristics related to this task. Throughout the literature, many studies are focused on the transcription of spoken language (BÉRARD *et al.*, 2018; NISHIKIMI *et al.*, 2019; SKOKI *et al.*, 2019). Despite sign language share many characteristics of spoken languages, only a few pieces of research bring automatic sign language annotation as the main factor of the study (MOCIALOV *et al.*, 2018; NAERT *et al.*, 2018; TAKAYAMA; TAKAHASHI, 2018).

In the same way as spoken languages, SLs emerged spontaneously, evolved naturally, and reflect the worldwide sociocultural differences, giving origin to a wide range of variations such as the British Sign Language (BSL), the American Sign Language (ASL), German Sign Language (DGS), the Chinese Sign Language (CSL), the Brazilian Sign Language (Libras), and many others. Due to the geographical extension and cultural diversity of Brazil, it is known that there are at least two recognized sign languages, namely: Libras and Urubu-Kapoor (FERREIRA-BRITO, 1986a). Besides, according to Silva and Kumada (2013), the context of deaf communities is, in fact, multilingual, welcoming several languages and linguistic varieties of a geographic order, age group, social class, influence of contact with oral and writing Portuguese, and borrowings from Portuguese or other SLs.

It is in the context of the Brazilian sign language that this work is inserted, and the main focus is the recognition of facial expressions using the Facial Action Coding System (FACS). To better present the work developed, the next sections present the research contextualization (Section 1.1), motivation (Section 1.3),the objectives (Section 1.4), the method adopted (Section 1.5), and the organization of this document (Section 1.6).

1.1 Research Context

Inserted in the problem of automatic transcription of facial expressions in Libras is the automatic recognition of facial expressions. Such an assignment appears in the literature heavily attached to emotion recognition, which leaves linguistic facial expressions to be studied separately by each sign language.

The task of automatic recognition of facial expressions in Libras has its first

work with Freitas *et al.* (2014) where was modeled a binary classification with Multilayer Perceptron Neural Network. Their focus was only GFE, and the framework displayed had as objective differentiate a grammatical from a neutral expression. That type of analysis allows studying the complexity involved in recognizing FE, although it demonstrated a weakness since it cannot differentiate between other GFEs. The dataset that supports this works is publicly available and has been successfully applied in various machine learning algorithms that boosted this binary approach (FREITAS, 2011; UDDIN, 2015; BHUVAN *et al.*, 2016; WALAWALKAR, 2017; XU, 2017; REZENDE *et al.*, 2016).

However, Libras has both affective and grammatical facial expressions embedded into them. So it is necessary a different and new approach. This work is the first to provide automatic analysis and classification of Libras' facial expressions associated with the facial action coding system. The present work represents a progression, as it brings a linguistic-based multiclass classification study, extensive experimentation with a combination of state-of-art deep networks, and it applies an automatic annotation method.

1.2 Contributions

The contributions of this study are:

1. A novel automatic recognition of facial action units;
2. Survey of facial expressions in Libras and its mapping into action units;
3. The building of two annotated video databases of Libras' facial expressions.

1.3 Motivation

Despite its proven linguistic potential, SL still represents a language of minority groups, and for that reason, it does not reach the prestige of some oral languages. Consequently, it does not receive the same incentive to make progress in research, professional training, production of teaching materials, etc. Therefore, without advancing in their first language (the SL), individuals with some degree of hearing impairment have difficulty in progressing in the oral or written Portuguese language. Such aspects frequently result in difficulties in the school and the work environment, preventing deaf people from reaching their full potential (SILVA *et al.*, 2020c).

Aiming to overcome the obstacles in the communication between hearing and deaf people, many efforts were dedicated, in the last decade, to the development of assistive technologies designed to promote greater inclusion of deaf people into the hearing society.

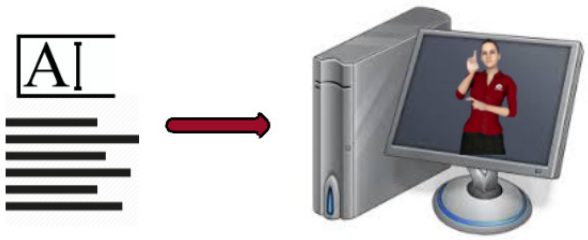


Figure 1.1 – Text-to-Sign Language (TTSL) technology translates text in a source language into a target sign language. The resulting system output is a virtual character animation (signing avatar) or a video of an interpreter.

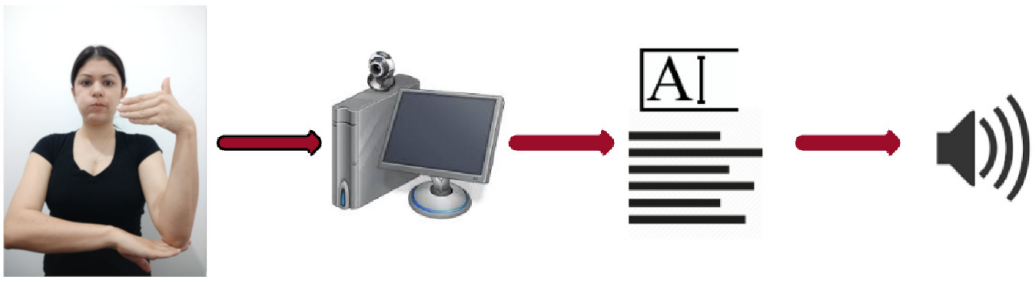


Figure 1.2 – Automatic Sign Language Recognition (ASLR) technology is designed to capture images, recognize and interpret sign language sentences performed by an individual, and transcribe them to text. Optionally, the text output can be the input of a Text-To-Speech (TTS) synthesizer, resulting in a translation from source sign language to target spoken language. ASLR systems have the potential to provide easier access to the deaf for public services or to guarantee private interactions with those that do not communicate in sign language without the help of an interpreter.

This is the case, for example, of the Text-To-Sign Language (TTSL) and the Automatic Sign Language Recognition (ASLR) technologies (Figures 1.1 and 1.2).

Automatic translation, synthesis, and recognition technologies for spoken languages are not new research topics. Indeed, such technologies have been experiencing a surprisingly fast pace evolution, boosted by the greater availability of digital samples of spoken and written content in multiple languages, the access to high-performance computing, and the recent advances in machine learning algorithms (in particular, deep learning techniques).

Still, a similar evolution was not observed for related SL technologies. The visuospatial domain of SLs poses significant challenges to the translation, transcription, and SL sentence segmentation.

Analogously to spoken languages, in which it is possible to combine phonemes to form words, SLs also present a set of basic meaningless segments that can be combined

to form meaningful signs. Liddell e Johnson (1989) define that segments in sign languages are composed of a set of *articulatory features*, that describe the *posture* of the hand, and the *activity* of the hand, encoded as *movements* and *holds*. Thus, unlike spoken languages, phonetic segmentation in sign languages requires not only the recognition of a set of articulatory features to identify a segment but also their sequential dynamics in time.

Another major challenge in developing SL communication technologies is the modeling of non-manual markers, which include facial expressions, head movements, and body movements. For numerous SLs, non-manual markers function as morphemes, and they also fulfill syntactic and pragmatic functions.

Finally, the computational modeling of any language depends on a structured and in-depth knowledge of the language and its grammar rules. However, most SLs worldwide can be considered understudied, meaning that aspects of their grammar and signs morphology are still undocumented or unknown. Besides, compared to spoken languages, there is a lack of annotated SL corpora, a key input for computational modeling.

1.4 Objective

In particular, we defined as the main objective of this thesis: to construct a robust video analysis framework capable of transcribing meaningful Libras' facial expressions using FACS. The detection and consequent recognition were solved as a pattern recognition problem modeled as multiclass classification.

Together with this primary goal, we also had the following specific objectives:

- To review the literature to identify significant facial expressions in Libras, under the supervision of a Libras specialist;
- To construct a comprehensive video corpus in Libras annotated using FACS (Facial Action Coding System);
- To propose an automatic facial expression recognition system for Libras;
- To develop an application that analyzes video of a signal interpreter and transcribes the recognized AUs to ELAN (Eudico Linguistic Annotator);
- To evaluate the proposed algorithm objectively, comparing the results of human and machine annotation.

We aimed at advancing and amplifying facial AUs recognition, contributing to the development of ASLR technologies but also providing a tool for linguistic studies

in Libras. We envision that the present work results could also be applied to other SLs and in affective computing, contributing to the recognition of facial expressions associated with emotions.

1.5 Method

Mathematical modeling is defined as describing the characteristics of a system using concepts and mathematical language to explain an occurrence, to describe or study the effects of different processes accurately, and to make predictions about behavior patterns. Our methodological approach for the mathematical modeling for the automatic annotation of facial expressions in sign language started with labeling the observable event, later moving towards acquiring data and then choosing the fittest recognition model based on the event we decided to detect and classify. To follow this logic, the methodology proposed for the present work is summarized as follows:

- review of the literature to identify the meaningful facial expressions in Libras, with the supervision of a specialist in Libras;
- definition of a transcription tool and proposition of an annotation model for Libras' facial expression;
- construction of a comprehensive video corpus in Libras annotated using FACS (Facial Action Coding System);
- proposal of an automatic facial expression AU recognition system for Libras;
- development of an application that analyzes video of a sign interpreter and transcribes the recognized AUs to ELAN (Eudico Linguistic Annotator);
- objective evaluation of the proposed algorithm comparing the results of human and machine annotation.

1.6 Document Organization

The present work is composed of seven chapters, being the first this introduction. Due to the interdisciplinary nature of our work, we start with a very detailed piece of fundamental concepts with machine learning basic notions and two chapters on facial expressions. The other chapters are occupied with methodology. The chapters are organized in the following way:

- Chapter 2 has basic concepts that we will use throughout this text and related work to facial expression recognition.
- Chapter 3 presents a bibliography study of Libras' facial expression and contextualizes its importance in the field.
- Chapter 4 presents a summary of Libras' transcription models, technical contributions, and limitations from developed studies.
- Chapter 5 describes the creation of the Libras' facial expressions dataset.
- Chapter 6 introduces our model for the resolution of recognition of Libras' FEs and the applied pre-processing step.
- Chapter 7 brings the description of experiment settings, results, and analysis resulting from the study.
- In the conclusion chapter, we present our final considerations, contributions, limitations, and future work.

1.7 List of Publications

During the elaboration of this thesis, the following works were published:

1. *SILFA: Sign Language Facial Action Database for the Development of Assistive Technologies for the Deaf.* Emely P. da Silva, Paula D. P. Costa, Kate M. Oliveira Kumada, and José M De Martino. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG) (pp. 382-386).
2. *Recognition of affective and grammatical facial expressions: a study for Brazilian sign language.* Emely P. da Silva, Paula D. P. Costa, Kate M. O. Kumada, José M De Martino, and Gabriela A. Florentino. In Sign Language Recognition, Translation & Production (SLRTP) Workshop at 16th European Conference on Computer Vision (ECCV 2020).
3. *Analysis of Facial Expressions in Brazilian Sign Language (Libras).* Emely P. da Silva, Kate M. Oliveira Kumada, and Paula D. P. Costa. In press: V Colóquio do Grupo de Pesquisa O corpo e a Imagem no Discurso: *Ceci n'est pas une pipe* e IV Simpósio em Transculturalidade, Linguagem e Educação: Thinking (and doing) otherwise. European Scientific Journal (ESJ).

2 Related Work

The analysis of human facial expression has been fascinating researchers for many centuries. In the nineteenth century, Darwin (1872) hypothesized that facial expressions must have had some instrumental purpose in evolutionary history.

In Computer Science, the origins of automatic facial expression recognition can be traced back to Face Recognition (FR) models, which involved the manual annotation of various landmarks on the face before presenting them to the computer for recognition (CHAN; BLEDSOE, 1965). In the early seventies, Ekman e Friesen (1971) published a cross-cultural study showing that, despite the vast cultural and language differences of interviewed subjects, a small set of six facial expressions were very frequently associated to the same emotions labels. The idea of universal facial expressions for happiness, sadness, anger, fear, disgust and surprise; guided first computer vision approaches to facial expression recognition.

In the following decades, the development of facial recognition technology was leveraged by the great interest of the industry in security and commercial applications. In the nineties, the evolution of FR systems was accelerated by the advent of digital image technologies and increased computational power and data storage capacities at a lower cost. Early FR systems, however, were too sensitive to variation in illumination, posing, and facial expressions. Facial expression recognition (FER) models emerged as a mechanism to improve face recognition accuracy. Many FR and FER systems adopt similar architectures and processing techniques (CHIBELUSHI; BOUREL, 2003).

In the last decade, significant advances in machine learning algorithms associated with greater computational power at a lower cost have boosted the results obtained by automatic facial expression recognition algorithms. In particular, the recent success of deep learning techniques in various fields has also been observed in the learning of discriminative representations for automatic FER (LI; DENG, 2020).

The first section of this chapter provides a brief introduction to computer vision algorithms based on neural networks. This section could be skipped by those who are experienced in machine learning algorithms.

Following, we present basic concepts and the general framework of automatic Facial Expression Recognition (FER) systems (Section 2.2). We also present an overview of the state of the art of deep facial expression recognition (Section 2.3). Following, in Section 2.4, we present the state of the art of FER algorithms applied to Automatic Sign Language Recognition (ASLR) technology. Finally, to conclude the chapter, we position

our work in the state of the art.

2.1 Neural Networks for Computer Vision: Basic Concepts

Artificial Neural Network (ANN) is a bio-inspired information processing paradigm. The idea is that in the way that our brain learns to accomplish tasks is possible for a machine to do the same. Thus, resorting to algorithms, neural network research applies mathematical and statistical foundation, concepts, and techniques. To model functions performed by the brain, the ANN constituent of processing units interconnected to ensure information storage and learning capacity. Due to their potential for learning and generalization, they are broadly applied to estimate solutions for complex problems that are otherwise intractable. A neural network has three essential elements: neural models, network architecture, and the learning process. In this work, we deal with the learning process, also called a learning algorithm or machine learning.

Learning in an ANN occurs through an adaptive process, where the weights in the processing units' connections are adjusted. From a set of data, representation algorithms transform feature information to predict the solution task automatically.

There are three types of learning: supervised learning, unsupervised learning, and reinforcement learning. The supervised learning process consists of the presence of input-output samples. The unsupervised learning method consists of a network learning by itself, i.e., the network parameters are optimized accordingly with a forecast measure of the assignment's quality. In reinforcement learning, the network performs its task several times and, upon completion, receives a grade for its result. This grade is an evaluation value of the performance of the task completed by the network (HAYKIN *et al.*, 2009; SILVA, 2016). In this work, we adopt supervised learning techniques.

2.1.1 Neural Networks for Feature Learning

Mathematically, a traditional supervised single-label classification associates an instance x with a single label y, from a finite previously known set of labels. In other words, there is a underlying function $f(x) = y$. The neural model also includes an externally applied bias, denoted by b. So, the resulting expression for an information processing unit becomes $f(x, b) = y$.

To make predictions of the target variable y, based on new values of x, one should find the best approximation of f that satisfies the equality. The problem then becomes $d_\varepsilon = d(\varepsilon) = d(f(x, b), y)$ where ε is a measure of the prediction error that we try to minimize. This can be done as an optimization problem, meaning that we compute

an objective function (also called functional risk $R[d_\varepsilon]$) that measures the error between the output scores and the desired patterns (HAYKIN *et al.*, 2009; LECUN *et al.*, 2015; SILVA, 2016).

Now, suppose that we are given a training set with N observations of x denoted $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$, together with corresponding observations of the labels $\mathbf{y}$ denoted $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$.

$$R[d_\varepsilon] = d(f(\mathbf{x} + b_N), \mathbf{y}) \tag{2.1}$$

We can solve this curve fitting problem by choosing the value of f for which $R[d_\varepsilon]$ is as small as possible (BISHOP *et al.*, 1995).

Since f can be very hard to estimate, we suppose that it is a polynomial approximation. So f can be described as $f(\mathbf{x}) = W\mathbf{x}$ where $W \in \mathbb{R}^{N \times N}$ is a matrix of adjustable weights. A machine learning algorithm then modifies its adjustable parameters to reduce the distance error by optimizing its weights. Note that, by exploring this probabilistic representation, in order to make predictions that are optimal according to appropriate criteria, our goal is the capability of generalization of a curve fitting algorithm to unseen data (BISHOP *et al.*, 1995; LECUN *et al.*, 2015; SILVA, 2016).

One simple choice for the functional risk widely used in the research community, is given by minimum square norm. So that we minimize

$$R[d_\varepsilon] = ||d_\varepsilon(\mathbf{y}, W\mathbf{x})||^2 = ||\mathbf{y} - W\mathbf{x}||^2 = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{N} W_{ij}x_j\right)^2, \tag{2.2}$$

where the constant $\frac{1}{2}$ is add for convenience. One solution for this problem is obtained by calculating $\mathbf{x}^* = (W^TW)^{-1}W^T\mathbf{y}$ and it is known as Optimum Linear Associative Memory (OLAM) introduced by Kohonen and Ruohonen in the 70s (KOHONEN, 2012).

As we can notice, there are various possibilities for a machine learning algorithm, given that there are many possible assumptions that could be made for the objective function and the distance measure.

Most applications use an iterative procedure called Stochastic Gradient Descendent (SGD) to find a solution to a differentiable objective function. This procedure has a good performance in finding appropriate weights (LECUN *et al.*, 2015). In practice, many image applications use a linear classifier on top of extracted features of some form. However, considering facial images, a linear or shallow classifier typically does not present robust results when dealing with pose or illumination variation. To improve the results, one can use non-linear features using kernels methods. With multiple non-linear layers, a system can learn to differentiate sensitive details. These are the significant advantage of the so-called Deep Learning models (LECUN *et al.*, 2015).

In the following sections, we present the most known learning machine techniques for deep network architecture, which are also adopted in this work. First, we define convolutional neural networks (CNN), which have consistently shown efficacy in recognizing facial expressions of emotion. Also, they compose robust architectures like the SqueezeNet. Following, by considering that facial expressions are a time observable occurrence, they can be explored with Recurrent Neural Networks (RNN), and we discuss the Long Short-Term Memory (LSTM) architecture.

2.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNN) exploits the spatial structure of data by learning local features. To be specific, the network is characterized by a partially connected feed-forward net. A general network layout has the source nodes in the input layer supplying respective elements of the input observation (activation pattern), which is applied to the computation nodes (neurons) in the second layer. The output signals from the second layer are the inputs to the third layer and so on for the rest of the network (HAYKIN *et al.*, 2009). A partially connected network is characterized by some communication links of the network not being connected. In a CNN, the top nodes have to satisfy a weight-sharing constraint, meaning that the same set of synaptic weights are used for each one of the nodes in the hidden layer. Then we may express the function f as follows:

$$f(x_j) = \sum_{i=1}^{N} W_i x_{i+j-1}, \quad j = 1, 2, \ldots, N, \tag{2.3}$$

which is in the form of a convolution sum. Besides the non-linearity hypothesis, the designed structure of such a network include the following forms of constraints (LECUN *et al.*, 1995):

Feature extraction: The extraction of local features is forced by the weight-sharing constraint (respective field) from the previous layer. Once a feature has been extracted, its position relative to other features is approximately preserved (HAYKIN *et al.*, 2009). The set of weights in the first few stages are called a filter bank (LECUN *et al.*, 2015).

Feature mapping: Since individuals neurons are constraint to share the same set of weights, each layer is composed of multiple feature maps. These feature maps take the form of a plan and by operating the convolution with a kernel or non-linearity such as ReLU[1], followed by a *sigmoid*[2] (squashing) function (HAYKIN *et al.*, 2009). Figure 2.1

[1] The rectifier is a function defined as the positive part of its argument: $f(x) = max(0, x)$, where $x \in \mathbb{R}$. A neuron employing the rectifier as its activation function is also called a Rectified Linear Unit (ReLU). In 2011, it was demonstrated that ReLU enables better training of deeper networks (GLOROT *et al.*, 2011).

[2] A sigmoid function is a family of functions that are characterized by an "S"-shaped curve or sigmoid curve. By definition, it is a bounded, differentiable, real function that is defined for all real input

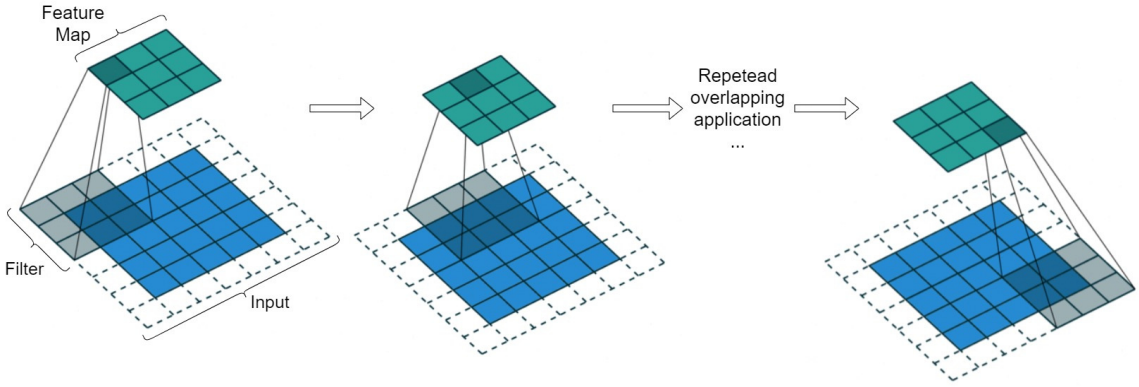


Figure 2.1 – Representation of the filter applied to an entry to create a feature map. This example is a convolution operation where the stride length is two. Note that the filter (kernel) is flipped before being applied to the input. As described, the convolution in the use of CNN can also be called a "cross-correlation" in Mathematics. Adapted from (SAHA, 2018).

brings an example of such convolution.

Subsampling or Pooling: After the convolutional layer, there is a subsampling computational layer to calculate the maximum of a local unit patch on one or a few resource maps. This operation reduces the sensitivity of the outputs to shifts, and other forms of distortion (HAYKIN *et al.*, 2009; LECUN *et al.*, 2015).

A bipyramidal effect is obtained with successive computational layers alternating between convolution and pooling (HAYKIN *et al.*, 2009), thereby reducing the dimension of the representation and creating an invariance. By backpropagating gradients through CNN, all the weights in all the filter banks are trained (LECUN *et al.*, 2015). In Figure 2.2 is presented such effect.

Classification: The last layers in a CNN are fully connected layers and that combines the features together by flattening the last output. Finally, an activation function such as *softmax*[3] or *sigmoid* can be used to classify the outputs. The whole CNN structure is presented in Figure 2.2.

Finally, in the feature mapping and the pooling stages, the network is abstracting different filters that, in practice, generate different feature maps from the same

values and has a non-negative derivative at each point (BISHOP *et al.*, 1995). Often, sigmoid function refers to the particular case of the logistic function and defined by the formula $\sigma(x) = \frac{1}{1+e^{-x}}$, given $x \in \mathbb{R}$.

[3] A softmax function is defined by taking as input a vector of M real numbers and normalizing it into a probability distribution consisting of M probabilities. So, independent of the entries value, after applying softmax, each component will be in the interval $(0, 1)$, and the elements will add up to one. In other words, the exit can be interpreted as probabilities (BISHOP *et al.*, 1995). Mathematically, The standard softmax function $\sigma : \mathbb{R}^M \to \mathbb{R}^M$ is defined by the formula $\sigma(\mathbf{v})_i = \frac{e^{v_i}}{\sum_{k=1}^{M} e^{v_k}}$, for $i = 1, \ldots, M$ and $\mathbf{v} = (v_1, \ldots, v_M) \in \mathbb{R}^M$.

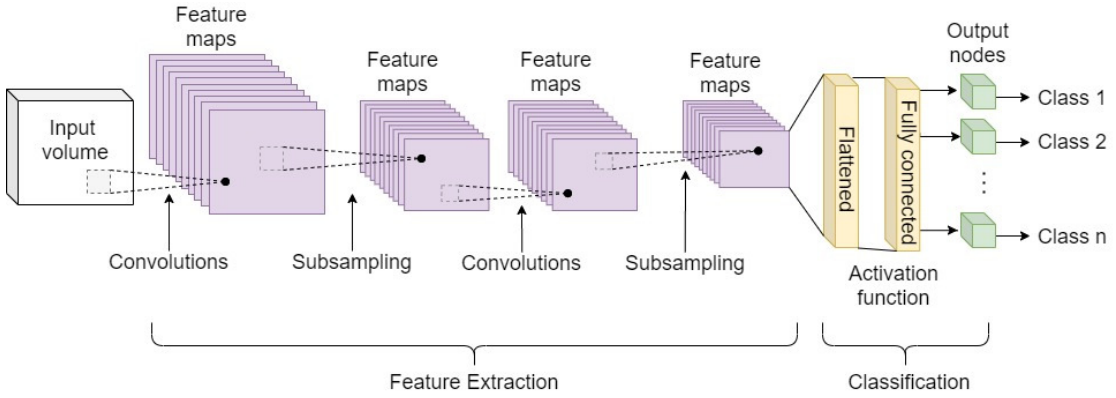


Figure 2.2 – A CNN sequence to classify an input volume. Source: The research itself.

original image. The increasing of the number of filters results in more image features that can get extracted. That may turn the network better at recognizing patterns in unseen images.

2.1.3 Long Short-Term Memory Networks

Long Short-Term Memory (LSTM), was introduced by Hochreiter e Schmidhuber (1997) and is a recurrent network architecture in conjunction with an appropriate gradient-based learning algorithm. LSTM applies to a range of tasks since it can process not only single data points but also entire sequences of data and was designed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional Recurrent Neural Networks (RNNs). This problem is particular to recurrent networks since difficulty with long-term dependencies arises from the exponentially smaller weights given to long-term interactions compared to short-term ones, i.e., the gradients are propagated over many stages.

Recurrent Neural Network: The architectural layout of a RNN takes many different forms. We consider a multilayer perceptron model where hidden neurons define the state of the network. For this theoretical RNN, we assume that the output is discrete and that the nonlinear function that characterizes the hidden layer is a hyperbolic tangent[4]. The hidden neurons define the state of the network. The production of each hidden layer is fed back to the layer via a bank of unit-time delays (HAYKIN *et al.*, 2009). The RNN internally computes the training loss with recurrent connections that maps an input sequence of x values to a corresponding sequence of output ϕ values. A natural way to represent discrete

[4] Hyperbolic functions are analogs of the ordinary trigonometric functions defined for the hyperbola rather than on the circle. They take a real argument called a hyperbolic angle, like instead of the points $(\cos t, \sin t)$ form a circle with a unit radius, the points $(\cosh t, \sinh t)$ form the right half of the equilateral hyperbola. The hyperbolic tangent is defined as $\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x}-1}{e^{2x}+1}$. Also, the hyperbolic tangent has two proprieties: is the solution to the differential equation $f' = 1 - f^2$ with $f(0) = 0$ and the nonlinear boundary value problem activation function (ZWILLINGER, 2002).

variables is to regard the output ϕ as giving the unnormalized log probabilities of each possible value of the discrete variable. We can then apply the softmax operation as a post-processing step to obtain a vector $\mathbf{y}$ of normalized probabilities over the output. Forward propagation begins with a specification of the initial state $\mathbf{h}^{(0)}$ (GOODFELLOW *et al.*, 2016). Then the general dynamic behavior of the RNN in response to an input vector, for each time step from $t = 1, ..., \tau$ is described by a system of coupled equations given as

$$\begin{aligned} \mathbf{a}^{(t)} &= \mathbf{b} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)}, \\ \mathbf{h}^{(t)} &= \tanh\left(\mathbf{a}^{(t)}\right), \\ \phi^{(t)} &= \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)}, \\ \mathbf{y}^{(t)} &= \sigma(\phi^{(t)}), \end{aligned} \tag{2.4}$$

where the parameters are the bias vectors $\mathbf{b}$ and $\mathbf{c}$ along with the weight matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, respectively, for input-to-hidden, hidden-to-output and hidden-to-hidden connections. The function σ is usually the softmax function. This is an example of a recurrent network that maps an input sequence to an output sequence of the same length (GOODFELLOW *et al.*, 2016). A full schematics of RNN operations can be followed in Figure 2.3.

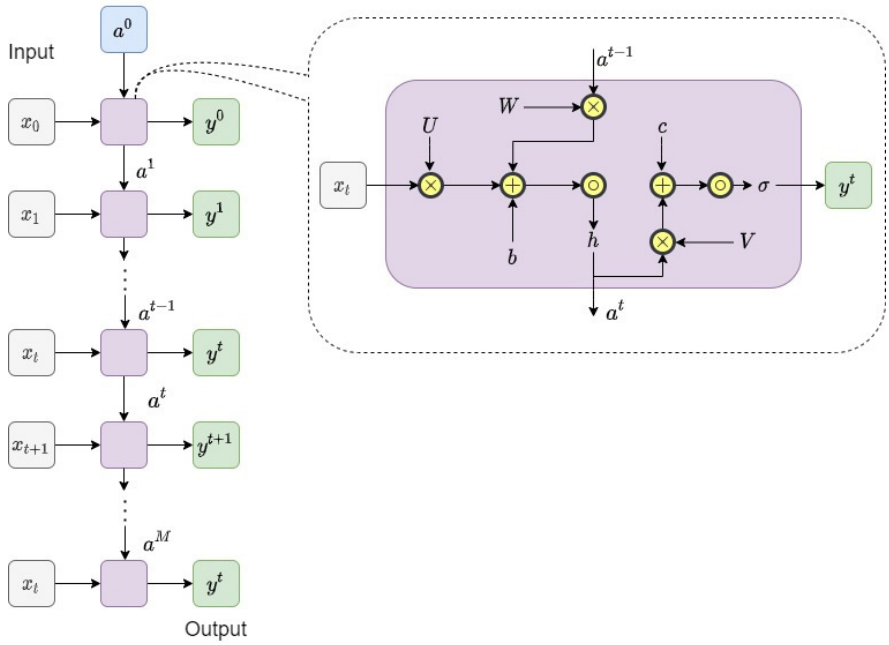


Figure 2.3 – A typical many-to-many RNN architecture. For each timestep t, as defined in the text, the entry is x^t, the activation state is a^t, the output is y^t. Also, W, U, V, b, c are coefficients that are shared temporally and h, σ activation functions. Source: The research itself.

LSTM. A very effective sequence model used in practical applications is a type of gated RNNs, called long short-term memory (LSTM). Gated RNNs generalize the connection weights that may change at each time step and are based on the idea of creating paths through time that have derivatives that neither vanish nor explode (GOODFELLOW *et*

al., 2016). Gated RNNs work by allowing the network to accumulate information and learning to decide when to clear the previous state.

In the long short-term memory model, self-loops conditioned in the context are created to produce paths where the gradient can flow for long duration. Accordingly with Goodfellow *et al.* (2016), besides the outer recurrence of the RNN where a unit simply applies an element-wise nonlinearity to the affine transformation of inputs and recurrent units, the LSTM recurrent networks have "LSTM cells" that have an internal recurrence (a self-loop). Each cell has the same inputs and outputs as an ordinary RNN but also has more parameters and a system of gating units that controls the flow of information. The most crucial component is the state unit $s_i^{(t)}$, which has a linear self-loop. A forget gate unit, $f_i^{(t)}$ for time step t and cell i, controls the self-loop weight (or the associated time constant) which sets this weight to a value between 0 and 1 via a sigmoid unit:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right), \tag{2.5}$$

where $\mathbf{x}^{(t)}$ is the current input vector and $\mathbf{h}^{(t)}$ is the currently hidden layer vector, containing the outputs of all the LSTM cells, and $\mathbf{b}^f$,$\mathbf{U}^f$,$\mathbf{W}^f$ are respectively biases, input weights, and recurrent weights for the forget gates. As follows, the LSTM cell internal state is thus updated but with a conditional self-loop weight $f(t)^i$:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right), \tag{2.6}$$

where $\mathbf{b}, \mathbf{U}$ and $\mathbf{W}$ respectively denote the biases, input weights, and recurrent weights into the LSTM cell. Similarly to the forget gate by using a sigmoid unit to obtain a gating value between 0 and 1, the external input gate unit $g_i^{(t)}$ is computed with its own parameters:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right), \tag{2.7}$$

Finally, the output gate $q_i^{(t)}$, which also uses a sigmoid unit for gating, can also shut off the output $h_i^{(t)}$ of the LSTM cell via:

$$\begin{aligned} h_i^{(t)} &= \tanh\left(s_i^{(t)}\right) q_i^{(t)}, \\ q_i^{(t)} &= \sigma\left(b_i^o + \textstyle\sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right), \end{aligned} \tag{2.8}$$

which has parameters $\mathbf{b}^o$,$\mathbf{U}^o$,$\mathbf{W}^o$ for its biases, input weights, and recurrent weights, respectively. Among the variants, one can choose to use the cell state $s_i^{(t)}$ as an extra input (with its weight) into the three gates of the i-th unit, which would require three additional parameters (GOODFELLOW *et al.*, 2016). Figure 2.4 brings a general LSTM schematics.

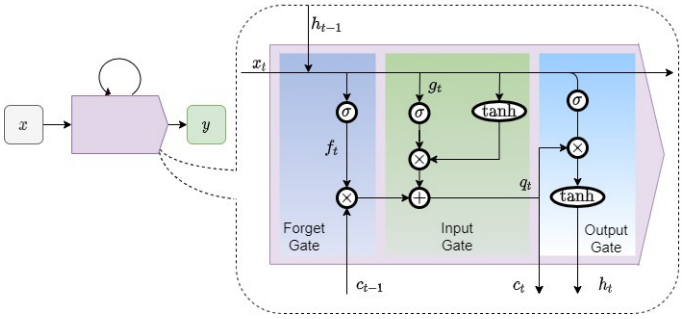


Figure 2.4 – Basic Structure of LSTM with the memory cell highlighted. LSTM module has three gates named as Forget gate, Input gate, Output gate. Further explanation can be found in the text. Source: The research itself.

Summing up, instead of having a single neural network layer, there are four interacting in the cell state, which is carefully regulated by gate structures that have the ability to remove or add information to the cell state.

2.1.4 Learning Process

As important as the network architecture is the learning process, or the training algorithm, which is an automatic method for estimating the parameters and updating the weights of a network based on training data. Implicitly, the learning process must decide which features of the input pattern should be represented by the hidden neurons (HAYKIN *et al.*, 2009).

A popular method type for the training is the back-propagation algorithm, which includes optimization algorithms like Stochastic Gradient Descent (SGD) (BOTTOU, 2010), and Adaptive Moment Estimation (ADAM) (KINGMA; BA, 2014) as special cases. The training proceeds in two phases: (1) forward phase, where the synaptic weights of the network are fixed and the input data is propagated through the network, layer by layer; (2) backward phase, in which an error is computed by comparing the output of the network with the desired response. In this second phase, successive adjustments are made to the synaptic weights of the network (HAYKIN *et al.*, 2009).

The development of the back-propagation algorithm provided a computationally efficient method for the training of NN. For further information and comparison studies about learning algorithms, see (SUTSKEVER, 2013; LECUN *et al.*, 2012).

2.2 Automatic Facial Expression Recognition: A General Framework

Many surveys on FER have been published in recent years (LI; DENG, 2020; SARIYANIDI *et al.*, 2015; FASEL; LUETTIN, 2003; PANTIC; ROTHKRANTZ, 2000), (ZENG *et al.*, 2009). From such works, a general framework for FER systems can be identified, which can be divided into four main components: image acquisition, preprocessing, facial expression modeling, and classification. Figure 2.5 brings a diagram for a general FER system.

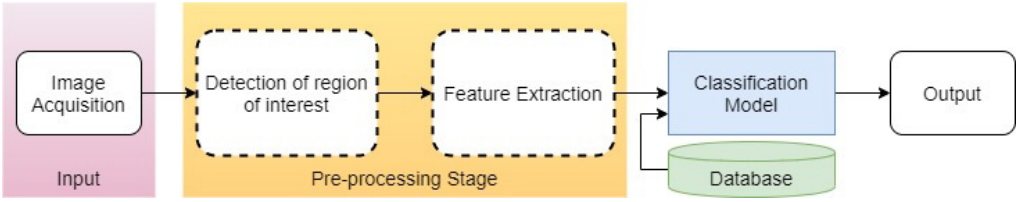


Figure 2.5 – Steps applied in a general FER system. Image created for this study.

2.2.1 Image Acquisition

The current standard in FER systems is the processing of color images (RGB), which are typically transformed into grayscale data in the early stages of the preprocessing step. Some recent works adopt depth images obtained from cameras with infrared sensors (LI; DENG, 2020). Depth cameras can associate depth information to facial pixels based on the distance from the camera, providing critical information of geometric facial relations. Depth information is naturally robust to pose and lighting variations but significantly increases the number of dimensions to be processed by any classification algorithm.

2.2.2 Preprocessing

Preprocessing in FER systems refers to all the steps that are performed after the image acquisition and before the face modeling. They are necessary because real-world scenarios may present multiple faces on an image, different backgrounds, occlusions, and variations in illumination and head pose (see Figure 2.6). The adoption of multiple data sources may also result in input images with different dimensions or resolution. Face detection, face tracking, image registration, image cropping, and some kind of normalization are common preprocessing tasks performed by systems to capture as much semantic meaning as possible from images.

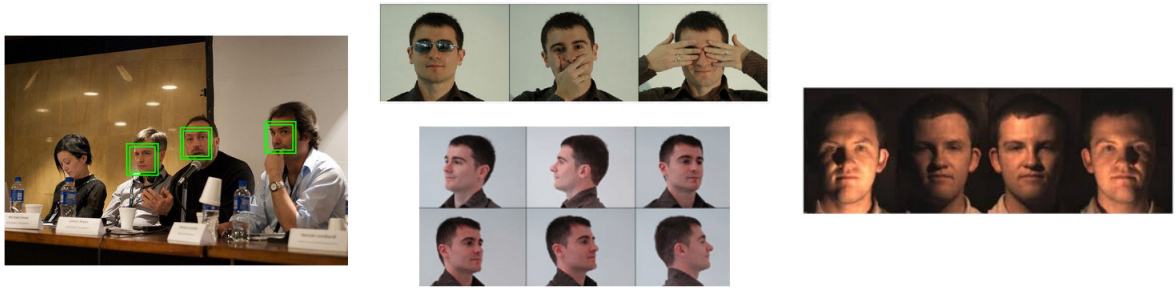


Figure 2.6 – After image acquisition, preprocessing steps are necessary to deal with common situations that are not ideal to perform FER. From left to right: multiple faces on a image, occlusion of important facial components, head pose and illumination variation. Adapted from (SHARIFARA *et al.*, 2014).

Despite still having challenges to be overcome, major advances have been made in preprocessing methods. This is the case, for example, of face detection and facial landmark tracking. Viola e Jones (2004) algorithm and the open-source software library derived from the work of Kazemi e Sullivan (2014), are classical and consolidated approaches that provide robust and accurate results in face detection and landmark tracking.

More recently, deep learning approaches to facial expression recognition, or deep-FER, introduced new steps into the preprocessing pipeline. Deep neural networks require sufficient training data to ensure generalization to a given recognition task. However, available training datasets usually do not have sufficient samples, and *data augmentation* steps are necessary. Most frequently, approaches generate new samples through the application of random perturbations and transformations such as rotation, shifting, skew, scaling, noise, contrast, and color jittering to the original dataset images (LI; DENG, 2020).

Other common preprocessing tasks and algorithms are broadly covered in the related literature (PANTIC; ROTHKRANTZ, 2000; FASEL; LUETTIN, 2003; ZENG *et al.*, 2009; SARIYANIDI *et al.*, 2015; LI; DENG, 2020).

2.2.3 Facial Expression Modeling

A key aspect in the characterization of FER systems is the set of adopted features to represent, or model, the face, and the facial expressions.

Sariyanidi *et al.* (2015) divide the existing approaches into systems that process individual input images and adopt frame-by-frame modeling (*spatial* or *static* representations) versus systems that take advantage of sequential frames in a video and process a range of input frames within a temporal window as a single entity (*spatio-temporal* or *dynamic* representations). Li e Deng (2020) highlight that *dynamic* representations are

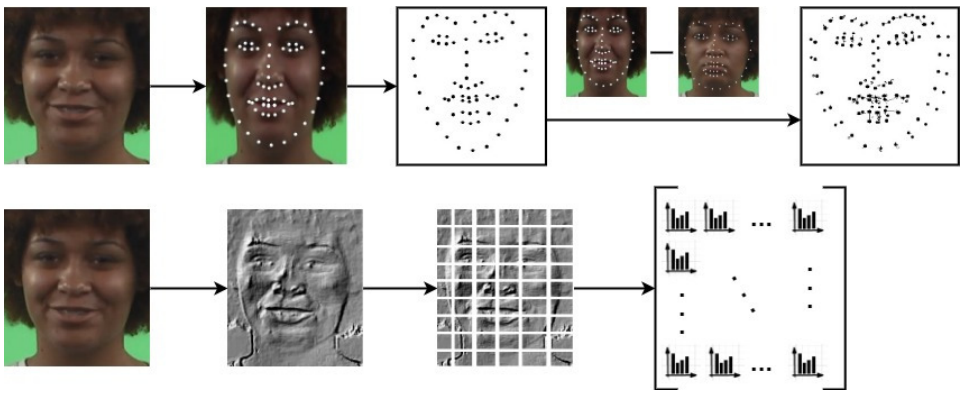


Figure 2.7 – Facial expressions can be encoded as *shape* (upper illustration), *appearance* features (lower illustration), or a combination of both. Features are transformed into a higher-level representation with the purpose of represent and generalize the data. The illustrations were inspired from (SARIYANIDI *et al.*, 2015), which provide a comprehensive review of feature extraction for facial expression analysis.

becoming a trend in deep-FER since they are capable of modeling temporal variation and subtle changes in facial expressions more efficiently. Comparing multiple deep-FER works, the authors show that training networks based on sequence data and analyzing temporal dependency can significantly improve the recognition performance on widely evaluated benchmarks.

A second categorization factor is a strategy adopted to encode the visual information of facial expression images. At this point, we observe a crucial difference between the first FER approaches and recent deep-FER systems.

Traditional non-deep facial expression recognition systems are characterized by adopting *handcrafted features* that are chosen based on the domain knowledge of the problem. Some works encode facial expressions as a set of *shape* or *geometrical* features, derived from the detection of key-points in the face (see Figure 2.7) (SHAN *et al.*, 2009). In other works, the *appearance* of the face is represented by some sort of texture encoding, like LBP (Local Binary Pattern) (CHENGETA; VIRIRI, 2018), HOG (Histogram of Gradients) (MALLICK, 2016), and Gabor filters (LYONS *et al.*, 1998a). According to Sariyanidi *et al.* (2015), high-level data-driven representations can also be derived from statistical encoding algorithms.

By contrast, a key characteristic of deep-FER approaches is that they typically process raw input images. Deep neural networks use multiple processing and filtering layers, implementing an *hierarchical* representation that progressively extracts relevant features, automatically *learned* from a large dataset of training images (SARIYANIDI *et al.*, 2015; BALNTAS *et al.*, 2016). In recent years, well-design network architectures have achieved state-of-the-art FER accuracy (LI; DENG, 2020; LIU *et al.*, 2018; ZHAO *et al.*,

2016b; PRAMERDORFER; KAMPEL, 2016).

Finally, existing approaches may also consider the face as whole unit (*holistic* representation) or as a set of parts (*analytic* or *part-based* representation). The latter ignores spatial relations among face parts and reduces the sensitivity to head-pose variation. Hybrid approaches that combine *shape* and *appearance* features or *holistic* and *part-based* representations, are also common (PANTIC; ROTHKRANTZ, 2000; SARIYANIDI *et al.*, 2015).

2.2.4 Classification

The final component of FER systems is the classification, or the recognition of facial expressions according to some categorization criteria.

From neuroscience and psychology, many works focus on the association between facial expression and emotions. However, facial expressions are used by humans to convey various types of meaning in different contexts. Different views about the problem result in different classification strategies.

Many FER systems adopt *affect* models and output an emotion label. Typically, a categorical model of emotions is adopted and the system analyzes the similarities of input facial expressions to archetypal facial expressions of a small set of emotions, such as the "big six" emotions of Ekman and Friesen (1971) (EKMAN; FRIESEN, 1971): happiness, sadness, anger, surprise, fear and disgust. In other works, the emotion labels are output according to a dimensional model such as the Pleasure-Arousal-Dominance (PAD) model of emotions (MEHRABIAN, 1996; COHEN, 2000).

As a strategy to recognize emotions, many works adopt the Facial Action Coding System (FACS) as the visual appearance building blocks of emotion. FACS refers to a taxonomy for a set of facial muscle movements that correspond to a displayed facial expression. The system is built upon Action Units (AUs), representing the muscular activity that produces momentary changes in facial appearance. Based on human anatomy, the FACS breaks down facial expressions into individual components of muscle movement. Created by Carl-Herman Hjortsjö with 23 facial motion units in 1970, FACS was further developed by Ekman e Friesen (1978).

The coding of facial expressions using FACS and the adoption of AUs as a measure of face variation gave rise to a whole set of FER systems in which the main classification stage is an attempt to recognize AUs, and later, in the post-processing stage, to classify those recognized AU into emotions (LIEN *et al.*, 1998; BARTLETT *et al.*, 2004; SIMPLICIO *et al.*, 2010; GHAYOUMI; BANSAL, 2016).

In the present work, we are not interested in reviewing FER systems that adopt a holistic facial expression emotion classification. Our study focuses on facial expressions that are not always related to the expression of emotions but also convey lexical and grammatical meaning in sign language communication.

In view of the above, it is believed that Action Units (AUs) are powerful facial expression descriptors and have shown their applicability beyond emotion analysis in areas such as language, mental health, and marketing. In the following section, we focus on reviewing the state of the art of AU-based FER systems.

2.3 State-of-the-Art Action Unit Recognition

Table 2.1 presents a summary of the key characteristics of the most relevant works that represent the state of the art of specialized AU-based FER systems. The following sections discuss the main aspects of their implementation.

Table 2.1 – Summarization of state-of-art in action units recognition

Machine Learning Technique	**Reference**	**Database**	**Features**	**#AUs**	**Model Architecture**	**Performance**
shallow	Bartlett *et al.* (2004)	CK	feature learning	18	SVM	MAR: 0.945
	Tong *et al.* (2007)	CK	feature learning	14	DBN	TPR: 0.87 FPR: 0.06 CR: 0.93
	Simon *et al.* (2010)	RU-FACS	feature learning	27	SO-SVM	AUC: 0.85 F_1 : 0.52
deep	Gudi *et al.* (2015)	BP4D+SEMAINE	learning based	14	CNN	F_1: 0.522 (11 AUs) F_1: 0.341 (6 AUs)
	Zhao *et al.* (2016b)	BP4D+	learning based	10	CNN	F_1:0.483 (12AUs) AUC: 0.56 (12AUs)
		DISFA				F_1:0.267 (8AUs) AUC:0.523 (8AUs)
	Walecki *et al.* (2017)	DISFA	learning based	14	CNN	ICC: 0.61(5AUs) MAE: 1.23(5 AUs)
		BP4D (FERA2015)				ICC: 0.45 (12 AUs) MAE:0.63 (12 AUs)
	Chu *et al.* (2017)	BP4D+GFT	learning based	12	CNN+LSTM	F_1:0.664 (GFT) F_1:0.825 (BP4D)
	Zhao *et al.* (2018)	EmotioNet	learning based	7	WSC	F_1:0.55
	Wang *et al.* (2018)	BP4D	learning based	17	LRBN	F_1: 0.66 (BP4D)
		SEMAINE				F_1: 0.66 (SEMAINE)
		CK+				F_1: 0.83 (CK+)
	Zhang *et al.* (2018)	DISFA+BP4D (FERA2015)	feature based	14	BORMIR	MAE: 0.852 (FERA2015) MAE:0.789 (DISFA)
	Liu *et al.* (2018)	DISFA+BP4D	learning based	12	GAM	MSE: 0.748 MAE:0.556
	Chu *et al.* (2019)	BP4D + GFT	learning based	12	CNN+LSTM	F_1: 0.664(GFT) F_1: 0.825 (BP4D)
	Deng *et al.* (2020)	Aff-Wild2	learning based	8	CNN CNN+RNN	Acc:0.938 F_1:0.236
	Kuhnke *et al.* (2020)	Aff-Wild2	feature based	8	3D ResNet	Acc:0.937 F_1:0.257
	PingAn-GammaLab	EmotioNet	-	11	-	Acc:0.9446 F_1:0.5659
	Ours	HM-Libras Silfa	feature based	119	CNN	F_1 : 0.74
					SqueezeNet	F_1 : 0.78

2.3.1 Action Units

Facial Action Coding System (FACS) is composed of action units, action descriptors, and visibility codes Ekman e Friesen (1978). The fundamental components are the *Action Units* (AUs), which are measurement units of muscular variation on the face. The constructors called *Action Descriptors* (AD) have no muscular basis for the action and no distinguishable specific behaviors. Also, *visibility codes* are labels for occlusion or unidentified action. For instance, when we smile, we use AU12. Also, when we show off the tongue, we use AU19 which is an action descriptor. Nevertheless, when we put a hand in front of the face, we have AU73, a visibility code. Apart from being visual and comprehensive, another advantage to this labeling is that AU can be observed in static or dynamics settings.

Action units can occur either singly or in combination (see Figure 2.8). While the number of action units is relatively small, it can be observable more than seven-thousand different AU combinations (HARRIGAN *et al.*, 2008). When they happen in combination, it can be **additive** or **non-additive**. If the combination does not change its constituents' appearance, they are additive, and if the shape does change, they are non-additive. Note that AU 1+2 is a non-additive combination since when AU2 occurs alone, it raises the outer brow and often also pulls up the inner brow, which results in a very similar appearance to AU 1+2. Forty-four AUs are defined as small distortions in the face that builds on complex facial expressions (EKMAN; FRIESEN, 1978). Thirty of them are anatomically related to the contractions of specific facial muscles on the upper face and eighteen on the lower face.

Despite the wide variety of AUs, the fifth column of Table 2.1 shows that most works are capable of recognizing just a small subset of action units. The ones related to categorical emotions are usually the chosen ones. Table 2.2 presents the action units

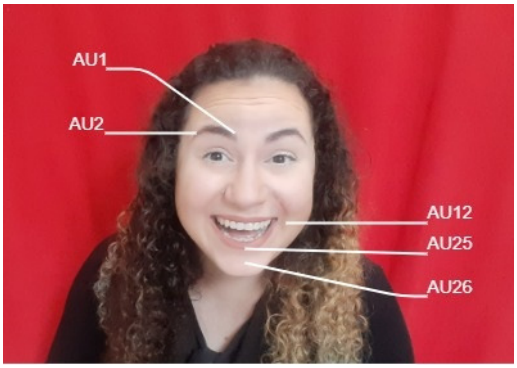


Figure 2.8 – Shown here are examples of AUs. The individual AUs displayed are 1, 2, 12, 25, 26. The AUs are combined as shown to produce the compound category; in this case, 1+2+12+25+26. Also, the labeled prototypical emotion category is happy, accordingly with Du *et al.* (2014).

Table 2.2 – Set of action units needed for prototypical emotions

Prototypical Expressions	**Involved Action Units**
Surprise	AU 1, 2, 5, 15, 16, 20, 26
Fear	AU 1, 2, 4, 5,15,20, 26
Disgust	AU 2, 4, 9, 15, 17
Anger	AU 2, 4, 7, 9,10, 20, 26
Happiness	AU 1, 6,12,14
Sadness	AU 1, 4,15, 23
Total number of AU involved	16

associated to the six emotions of Ekman (EKMAN, 1992; EKMAN, 1993). Works that classify combinations of AUs are rare (SIMPLICIO *et al.*, 2010), and works that combine the recognition of facial AUs with head pose AUs are nonexistent (TONG *et al.*, 2010). Those works that take into account combinations of action units, in general, suffer because of the modeling and datasets of the emotions' facial expressions being focused on the six basic emotions. Thus, recognition becomes rigid, and the FE is only classified if the exact combination of action units is found in the images. Any variation depends on a different approach. While studies that consider head movement presents an even greater challenge since they must perform head tracking in conjunction with the classification of facial expressions. It is very tricky to extract different feature sets from the face and then combining them. So, not many researchers choose to adventure into these methods.

2.3.2 Databases for AU Recognition

In the second column of Table 2.1, we observe that most AU-based FER systems adopt annotated databases for supervised learning (learning by example). However, FACS annotation is a time-demanding task and requires extensive training. For this reason, the majority of annotated image databases mentioned in the FER literature are encoded with emotion labels and just a few of them — all of them also created for emotion studies — are labeled with AUs. Japanese Female Facial Expression (JAFFE) Database (LYONS *et al.*, 1998b), Karolinska Directed Emotional Faces (KDEF) Database (LUNDQVIST *et al.*, 1998), The Belfast Induced Natural Emotion Database (BINED), Multimedia Understanding Group (MUG) Database (AIFANTI *et al.*, 2010), Radboud Faces Database (RaFD) (LANGNER *et al.*, 2010), Oulu-CASIA NIR-VIS database (ZHAO *et al.*, 2011) are only annotated with emotion labels, and a few of them also provided their intensity. In the computer vision field, there are some smaller databases which are annotated with a subset of emotions, and a subset of AUs like pose variations and smile. For instance, AR Face Database (MARTINEZ, 1998), Yale Face Database (GEORGHIADES *et al.*, 1997), Indian Spontaneous Expression Database (ISED) (HAPPY *et al.*, 2017), CVL Face Database, FEI Face Database and, the MPLab

GENKI Database.

However, all these databases are FACS coded. In other words, they are not readily available for training AU detectors, and they lack samples of common sign language facial expressions. Among the databases that include AU annotation, we highlight CK and DISFA (KANADE *et al.*, 2000; MAVADATI *et al.*, 2013; ZHANG *et al.*, 2014).

Cohn-Kanade AU coded expression database has the first release called CK, which includes 486 sequences from 97 subjects posing the six basic Ekman's emotions (KANADE *et al.*, 2000). Each sequence starts with neutral and ends in an apex of emotion and is AU coded. The second release is called CK+ and includes both posed and non-posed expressions (LUCEY *et al.*, 2010). Validated emotion labels have also been added to the metadata. In addition, CK+ provides baseline results for facial tracking, AU and emotion recognition. It is important to remark that the AU annotations were given at video and not frame-wise. The database is available[5] for research purposes and non-commercial use.

Denver Intensity of Spontaneous Facial Action (DISFA) database is a spontaneous database composed of videos of 27 subjects (12 females and 15 males) that vary in age from 18 to 50 years (MAVADATI *et al.*, 2013). The subjects are filmed while reacting to an emotional four-minute video stimulus. Also, it comprehends the manually labeled frame-based annotations of 5-level intensity of twelve FACs (1,2,4,5,6,9,12,15,17,20,25,26), labeled by two FACs coders. Each video has 4845 frames acquired at 20 fps, giving a total of 130.815 images. In addition, it is included the 66 facial landmarks points of each image. The lack of available data for comparing posed and spontaneous expression encouraged the same research group to construct the Extended DISFA Dataset (DISFA+) (MAVADATI *et al.*, 2016), which contains the videos and AU annotations of posed and spontaneous facial expressions of 9 participants in the same format as DISFA. DISFA+ also provides ground-truthed data, landmark points, subject-based self-report, and quantitative and qualitative comparison between posed and genuine facial muscle activations (MAVADATI *et al.*, 2016). The DISFA database is available[6] for use in research purpose upon an agreement request.

2.3.3 AU Recognition Models

The fourth and the sixth columns of Table 2.1 show the different modeling architectures adopted by AU-based FER systems.

For feature extraction, various methods are designed to capture facial geometry and appearance features caused by the variation of each action unit. Given that AU

[5] <http://www.pitt.edu/~emotion/ck-spread.htm>
[6] <http://www.engr.du.edu/mmahoor/DISFA.htm>

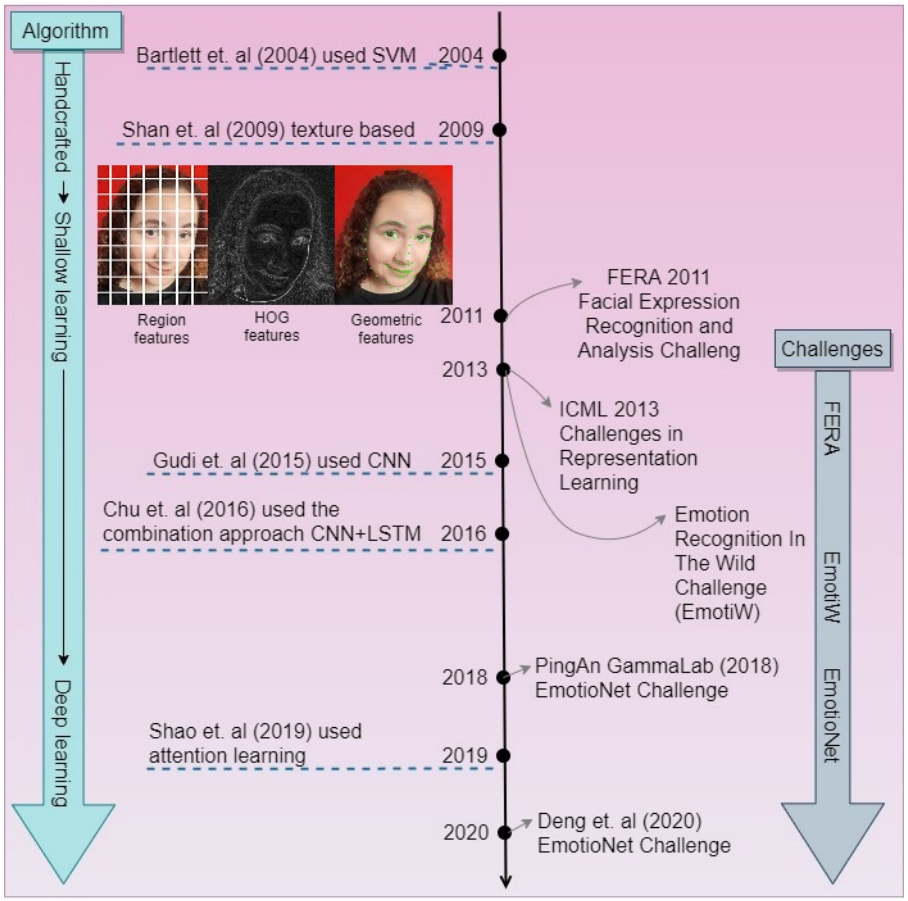


Figure 2.9 – The evolution of facial expression recognition techniques and challenges. Adapted from (LI; DENG, 2020).

occurs in sparse space on the face, some algorithms employ *Region Learning* (RL). Such a technique aims to identify and select specific regions to improve detection accuracy (ZHAO *et al.*, 2018). Head movement and occlusions is a problematic factor that affects the region separation, because fixed regions do not follow the action of the face, staying at the same place. So learning the parts aims to overcome this problem making the network focus on the desired region at the same time, and is applied by Zhao *et al.* (2016b) and Batista *et al.* (2017). In a way to diminish the influence of uncorrelated pieces of information, sparsity-induced[7] can be united with crop or texture to highlight the desired regions (ZHAO *et al.*, 2016b). Geometric measures can also be used as features together with texture or alone for AU classification (GHAYOUMI; BANSAL, 2016; BENITEZ-QUIROZ *et al.*, 2017). In Figure 2.9, we showed the evolution of AU detection models throughout the years and examples of handcrafted features.

[7] By sparsity-induced, we refer to the property that a subset of the model parameters have a value of precisely zero. Sparsity is also commonly used to refer to the proportion of a neural network's zero-valued weights. Higher sparsity corresponds to fewer weights and smaller computational and storage requirements. With zero-valued weights, any multiplications (which dominate neural network computation) can be skipped, and models can be stored and transmitted compactly using sparse matrix formats (GALE *et al.*, 2019).

Some researchers consider AU correlations where the presence of a specific AU could increase or decrease the likelihood of other AU. This assumption is closer to real-life occurrences. Algorithms with multi-label training[8] improve the detection by assuming that exists a correlation between labels. Also, studies confirm that multi-label strategy has potential for addressing unbalanced data (ZHAO *et al.*, 2018; ZHAO *et al.*, 2016b). To describe AU correlations Bayesian Networks[9] (SIMPLICIO *et al.*, 2010; SINGH *et al.*, 2014), dynamic Bayesian Networks (TONG *et al.*, 2010), or deriving a statistic correlation (CHU *et al.*, 2017; HAMM *et al.*, 2011) as a preprocessing stage has proved to have acceptable performance (WANG *et al.*, 2018).

Deep-FER approaches are also highlighted in the second column of Table 2.1. The most popular modeling approach is convolutional neural networks (CNN). Liu *et al.* (2013) introduced a deep network with AU-aware receptive field layer, designed to search subsets of the over-complete representation, each of which aims at best simulating the combination of AUs. Then, its output is passed through additional layers aimed at the expression classification, showing a large improvement over the traditional handcrafted image features. A seven-layer CNN for estimating AU occurrence and intensity is used by Gudi *et al.* (2015). Ghosh *et al.* (2015) adopted a multi-label CNN to show that a shared representation can be directly learned from input images. Using cascade regression framework to captured global AU relationships and global dependencies between AUs and landmarks, Wu e Ji (2016) demonstrate that the intertwined relationship of facial action units and face shapes boost the performances of both facial landmark detection and facial action unit recognition. A deep region multi-label learning (DRML) was proposed by Zhao *et al.* (2016b), with an intermediate region layer that is able to learn region-specific weights of CNNs. For each input image, the region layer returns an importance map and the network is trained for joint AU detection. A different approach was introduced by Li *et al.* (2017b) where for each facial region, an individual CNN is designed through facial regions cropping around the detected landmarks, called Enhancing and Cropping (EAC) net, where E-net enhances region of interest (ROIs) of the facial landmark features with attention map and c-net. However, this approach has been criticized for having limited generalization ability when applied to new subjects or new AUs, which usually have data distributions different from the training set (LEE *et al.*, 2019). All these methods focus solely on feature extraction while the network output remains unstructured. But Walecki *et al.* (2017) capture the output structure employing a conditional random field

[8] Multi-label classification is defined as the algorithm that assigns to each sample a set of target labels (ZHANG; ZHOU, 2014).

[9] Bayesian networks are a type of statistical model that represents a set of probabilistic relationships, i.e., they are a type of probabilistic graphical model that uses Bayesian inference for probability computations. For more information, see Neapolitan *et al.* (2004).

(CRF)[10] graph for AU dependencies that explicitly account for ordinal and non-linear relations between multiple results directly from various datasets. Generative adversarial recognition network (GARN) that adapts the model from the source subjects to a target subject for personalized AU recognition in an unsupervised way is proposed by Wang e Wang (2018). Shao *et al.* (2019) present an attention learning model where spatial attentions are adaptively learned using only AU labels.

Some models adopt a *spatio-temporal* or *dynamic* modeling (see Section 2.2.3). Such works consider the different facial expression phases (onset, apex and, offset) to establish a temporal manifold for AU and a sliding window to produce representations (LIU *et al.*, 2018; WALECKI *et al.*, 2017; ZHAO *et al.*, 2016a; GUDI *et al.*, 2015). Incorporating temporal modeling, recurrent neural networks are chosen to combine with CNN in Donahue *et al.* (2015). Jaiswal e Valstar (2016) proved that using temporal information throughout an LSTM (Long Short Term Memory) method could improve the performance of facial AU classification by proposing a CNN+LSTM model, which can learn the geometric, appearance, and dynamics information jointly for AU detection. In Wu *et al.* (2015), both static frames and motion optical flow, combined with two CNNs and LSTMs, were fused to produce a per-video prediction. Similarly, Chu *et al.* (2017) stacked LSTMs on top of a fusion network that receives the output of a CNN trained for recognition of facial spatial features. The AU detection performance in this work has improved significantly compared to existing approaches, due to the fusion of both spatial CNN and temporal features. A deep learning framework for AU detection with the region of interest adaptation and multi-label learning that also uses LSTM to utilize temporal information was proposed by Li *et al.* (2017a). Such models considered that AUs are active in sparse facial regions, and RL or ROI just aims to identify these regions for better specificity. Ma *et al.* (2019) introduced a unified end-to-end learning model that encodes prior expert knowledge, called AU R-CNN. They used various dynamic models (including two-stream network and CNN+LSTM) incorporated into AU R-CNN, and they affirm that such temporal dependency cannot always improve performance in all cases. Column "Model Architecture" of Table 2.1 summarizes this discussion.

With the raising in AU detection approaches, the need for a fair comparison of results was necessary. A way to level the plain field was creating open challenges, where a group of researchers provides a task for recognition and an annotated database in a time frame. Challenges in facial expression recognition started in 2011 with the Facial Expression Recognition, and Analysis Challenge (FERA)(VALSTAR *et al.*, 2011). In 2017, the FERA challenge (VALSTAR *et al.*, 2017) focused on solving AU detection, and in-

[10] Conditional random field (CRF) are a class of statistical modeling method used for structured prediction. In contrast, a classifier predicts single sample label without considering "neighboring" samples; a CRF can take context into account.

tensity estimation under various head poses; the winners Tang *et al.* (2017), Zhou *et al.* (2017) adopted CNN based approaches. EmotioNet challenge imposes the task of detecting and annotate AUs through recognition of AU and intensity; the winner in 2017 was the PingAn-GammaLab group network (BENITEZ-QUIROZ *et al.*, 2017). Affective Behavior Analysis in-the-wild (ABAW) (KOLLIAS *et al.*, 2020) competition in 2020 had an action unit detection track. Deng *et al.* (2020) has got the first position and became state of the art in AU detection. They designed a generic architecture for multitask learning in a teacher-student model, and they highlight the importance of data balancing for classification tasks in multitask learning. For the eight FACS AUs in the Aff-Wild2 database (KOLLIAS; ZAFEIRIOU, 2019), the detection achieved an F1 score (DERCZYNSKI, 2016) of 0.236 and average accuracy of 0.938 . In the second place, Kuhnke *et al.* (2020) designed a 3D ResNet (TRAN *et al.*, 2018) with face-alignment to guide the model to learn person independent face-region-related features. The published detection achieved an F1 score of 0.257 and average accuracy of 0.937. In Table 2.1 is possible to compare these results with some approaches mentioned above.

Recent methods such as (BALNTAS *et al.*, 2016; SONG *et al.*, 2016; SCHROFF *et al.*, 2015; FERNANDEZ *et al.*, 2020; ZHANG; GU, 2020) approaches the problem of learning AUs class by the adoption of a particular cost function, which in learning methods imposes constraints on the solution space, whose shape can take any form satisfying the underlying properties induced by the chosen loss function. The idea is that the optimization of the cost function leads to creating a solution space where every object has the nearest neighbors within the same class.

Finally, it is essential to remember that we have vast training data, but relatively little AU labeled training data. Some works try to overcome such an obstacle with semi-supervised and unsupervised learning strategies (LEE *et al.*, 2019; LI *et al.*, 2019). Unsupervised learning uses unannotated data and can be broadly categorized into semi-supervised and weakly supervised learning. In semi-supervised learning, the algorithm uses labeled data to extend annotation to a similar cluster or follow the unannotated data's continuity. Also, weakly supervised learning exploits the weak annotations; in other words, it builds on incomplete, inaccurate, inexact labeling. Preserving similar appearance and semantics, a weakly supervised model optimizes and extends learned features. Weakly Supervised Clustering (WSC) can prune noisy annotation and is adopted by Zhao *et al.* (2018).

2.4 State of the art of FER for Sign Languages

Sign language recognition is a growing topic around the world. Automatic Sign Language Recognition (ASLR) researchers have two ways to categorize spatio-temporal sign recognition: isolated and continuous recognition (KELLY *et al.*, 2011). The focus in isolated recognition is the prediction for one static gesture performed at a given time. In comparison, continuous recognition has an object of a sequence of gestures, where the aim can be to spot, segment, and classify meaningful articulators from within a signed sentence (KELLY *et al.*, 2011; IBRAHIM *et al.*, 2020).

Ong e Ranganath (2005) highlight that non-manual features have not received attention in the literature, despite the relevance of the information in a sign conveyed through this non-manual channel. Frequently, signs that are identical regarding the manual features are only distinguishable by the non-manual features accompanying the sign. There is difficulty identifying exactly which elements are essential to the sign and which elements are coincidental. In the last years, there has been some work towards the facial features, and how to combine the information from the manual and non-manual streams (COOPER *et al.*, 2011). Fundamentally Aran *et al.* (2009) identified a two-step classification process to integrate manual and non-manual features in an ASLR. When there was ambiguity, they introduced a second stage classifier to use non-manual expressions to resolve the problem. While this might appear a viable approach, it is not clear from sign language linguistics that it is scalable and generalized into other SLs (COOPER *et al.*, 2011).

On the other hand, Agris *et al.* (2008) used the Active Appearance Model (AAM) to merged manual and non-manual markers in the recognition process. Following this approach, they showed that some DGS signs could be recognized based on non-manual features alone. Generally, the recognition rate increases by between 1.5% and 6% upon inclusion of non-manual features.

The task of recognition is often simplified by forcing on recognizing isolated instances of manual or non-manual markers. In the manual case, hand position used in alphabets and numbers, instead of gestures from words and sentences composed with movement transitions (COOPER *et al.*, 2011; IBRAHIM *et al.*, 2020). Existing works in automatic sign language facial expression recognition can take a part-based or a holistic approach (see Section 2.2.3)(CARIDAKIS *et al.*, 2014).

As in the case of the lips, researches in pattern recognition for Lip-Reading combined with manual articulators may facilitate ASLR models because the production of visual syllables with the mouth while signing is part of the sign language (called mouthings, visemes, or spoken word). Some amount of work has been done in analysis of

mouthings for Sign Language of the Netherlands (NGT)(BANK *et al.*, 2011; BANK *et al.*, 2015), Irish Sign Language (MOHR, 2012), and German Sign Language (SCHMIDT *et al.*, 2013; ANTONAKOS *et al.*, 2015; KOLLER *et al.*, 2015). On the other hand, lip shape recognition is a well-established field because it is also applied in lip reading for security purposes and speech recognition (COOPER *et al.*, 2011). Approaches like (LAN *et al.*, 2009) uses an Active Appearance Model to track the lips, while (ONG; BOWDEN, 2008) recognize phonemes from the lips by extending to include HMMs has obtained significant attraction. For more information on lip-shape recognition, see Antonakos *et al.* (2015), Borysova (2017). In Heracleous *et al.* (2009), the handshapes from cued speech[11] was used to improve the recognition rate of lip-reading significantly. They model the lip by using some basic shape parameters. Also, by taking hand gestures from cued speech as disambiguate vowels in spoken words for lip readers. An early work combining manual sign recognition with viseme shapes in signs was Koller *et al.* (2014), where it is used Expectation-Maximization (EM) with Gaussian clustering in an HMM-framework to train viseme models from GSD corpus. More recently, Koller *et al.* (2019) split the recognition task up into sub-problems that occur in parallel. Sign-gloss, mouth shape, and hand shape classifiers are treated with embed CNN-LSTM models in each HMM stream following a hybrid approach at the GSD signs end.

The eye gaze is indispensable in the SL dialogue, and it can indicate a person and places. In Libras Felipe (1997) describes that even unaccompanied by the use of hands, the look and a slight movement of the head can be used as a deictic. Existent works on the subject of gaze gestures, in terms of possibilities (DREWES; SCHMIDT, 2007; DREWES, 2010), analyze the impact of different temporal codifications on performance (ROZADO *et al.*, 2011), and limitations of the technology (ROZADO *et al.*, 2012).

The problem of head motion recognition can also be described as pose estimation and have been obtaining better recognition rates with the usage of deep neural networks, and multimodal approaches (PIGOU *et al.*, 2017; PIGOU *et al.*, 2018; MURPHY-CHUTORIAN; TRIVEDI, 2009). Combining multi-scale and spatial-temporal analysis for small sign language units, Liu *et al.* (2014) proposed a system to recognize head and eyebrow movement in ASL.

Benitez-Quiroz *et al.* (2014) treats the features defining ASL non manuals gestures as a whole. They introduce a study correlating head movement, facial expressions, and linguistic features. Nguyen e Ranganath (2008a) uses Lucas-Kanade-Tomasi feature tracker to analyze the difficulties posed by inter-signer differences and using cluster from

[11] Cued Speech is a visual system used for communication between deaf and hearing people. It is a visual mode of communication that uses handshapes and placements combined with the mouth movements and speech to make the phonemes of spoken language look different from each other (ASSOCIATION *et al.*, 2006).

probabilistic PCA to overcome them. In more recent work, to recognize four sign language facial expressions, they combine HMMs, and Neural Networks (NN) (NGUYEN; RANGANATH, 2008b). As a way to improve tracking of facial expression during occlusion by the hands, (VOGLER; GOLDENSTEIN, 2005; VOGLER; GOLDENSTEIN, 2008; VOGLER *et al.*, 2007; KRINIDIS *et al.*, 2009) used a deformable surface model to track the face and later learn to classify their respective model by a learning algorithm (Radial Basis Function Networks and Boosted Input Selection Algorithm for regression) (BA; ODOBEZ, 2008; BAILLY; MILGRAM, 2009).

Ming e Ranganath (2002) choose to separate affective and grammatical facial expressions. They split the face into lower and upper face channels then performed Independent Component Analysis (ICA) on PCA for training data. Later then compare with a representation for Gabor wavelets networks (MING; RANGANATH, 2002).

To date, there are six works (FREITAS, 2011; UDDIN, 2015; BHUVAN *et al.*, 2016; WALAWALKAR, 2017; XU, 2017; REZENDE *et al.*, 2016) that treat the recognition of grammatical facial expressions in Libras and are considered state-of-the-art. Five of them (FREITAS, 2011; UDDIN, 2015; BHUVAN *et al.*, 2016; WALAWALKAR, 2017; XU, 2017) are due to the same data repository 'Grammatical Facial Expression Dataset', created by Freitas (2011). We point out that they called grammatical facial expressions only expressions that define a type of sentence, excluding expressions that are fundamental to describe the intensity or differentiate a sign from another. In Rezende *et al.* (2016) the proposed model has been only tested on the same database, being sign dependent constraint. As in any recognition task, subject independence[12] is vital to be capable of achieving high recognition rates and of being according to real-world conditions. With this type of structure, it becomes challenging to implement such an approach with other systems.

2.5 Concluding Remarks

In this chapter, we introduced basic concepts and a brief explanation of the framework for facial expression recognition methodologies. Along in the chapter, we showed the evolution of action unit detection and the distinct approaches available in the field historically. As this work is also inserted in recognition of sign language's facial expressions, we mentioned works in the area from different sign languages.

Table 2.1 positions our work. From the table we highlight two contributions of the present work: (1) we built a new AU-annotated dataset; (2) we implement a classifi-

[12] Subject Independence is a capability of an algorithm where it can generalize learning throw new subjects on unseeing data.

cation model trained with a wide set of action units.

Considering the state of the art of FER in Sign Language systems, we also bring a different broader view to SLs facial expressions by using FACS and experiments with a deep learning approach.

The following chapters present our approach to the modeling of a FER system for Libras' facial expressions using FACS.

3 Facial Expressions in Libras

As a first contribution of our work and a first step in our methodology, we conducted a survey to identify the elementary characteristics of facial expressions and head movements that are associated to affective and grammatical facial expressions in Libras.

In Brazil, Libras was recognized by the legal system as a linguistic system in 2002 (BRAZIL, 2002), and it is the linguistic system for transmitting ideas and facts from deaf communities in Brazil. The studies on Libras grammar increased relevance after the official language status and, despite linguistic studies on Libras can be found since the eighties (FERREIRA-BRITO, 1986b), there is still a small number of works dedicated to the study of facial expressions in Libras (SOUZA *et al.*, 2010; SOUZA, 2014).

Works on sign language show a variety of terminology given to the set of facial expressions applied in sign language according to their linguistic function. Baker (1985) used the term *nonmanuals* to describe facial behavior that carries grammatical, lexical and modifies adverbs and adjectives when co-occurring with manual signs. Benitez-Quiroz (2015) has used the same term nonmanuals to describe the facial features that define the constructors of dynamical facial expressions of grammar. Both works treat sign language in general; however, their analysis is done only for American sign language. For Libras, Paiva *et al.* (2018) included the parameters facial expression, head movement, and body movement as *non-manual expressions* (NME) to describe Libras phonologically by following the works of Ferreira-Brito (1990), Ferreira-Brito (1995), Baker e Cokely (1980). Also accordingly with Rezende *et al.* (2016) and Almeida (2014), non-manual expressions are features that can qualify a sign and add to its meaning, as well as being specific identifiers of a given sign (ALMEIDA, 2014).

Alternatively, Arrotéia (2005) defines the movements performed by the head, torso, shoulders, and facial expressions involving eyebrows, eyes, nose, mouth, and cheeks as corresponding to *non-manual markers* (NMM). Thus, a non-manual mark is a description of visual face and body features. This non-manual marking co-occurs with the manual signs in the sentences. The terms nonmanuals, non-manual expression, and non-manual marker can appear synonyms. Still, as we had stated, they define different studies of facial expression, head movement, and body motion in sign language. Due to non-manual markers be more precise and comprehensive in SL facial expressions, for this work, we will assume the non-manual markers nomenclature to describe facial expressions and head movements in both grammatical and affective applications. We highlight that despite the

fact that non-manual markers are related to the movement of the head, face, and body (or torso), in this work we focus on facial expressions only.

We found necessary to compile a survey of references dedicated to the description of Libras, with particular attention to those references that described the production of non-manual markers. The definition of the relevant keywords and references for this study was established through interactions with deaf individuals, sign interpreters and linguistic researchers. The guidelines adopted in the study are described in Section 3.1.

As a result of our study, Section 3.2 presents an overview of facial expressions in Libras, describing their function in the discourse, as previously said, excluding non-manual features that has an association with body movement.

Finally, in Section 3.3, we compare facial expressions detailed by various authors and we propose a compiled taxonomy that is adopted in our classification methodology.

3.1 Survey on Libras' Facial Expressions

The carried survey of the production of works in Libras that deal with facial expressions was started by analyzing books and other references indicated by professionals in the field. Our first contact with an examination of non-manual markers was through the advisers of this work, which are part of the assistive technology for the Deaf (TAS, Brazilian Portuguese acronym for *Tecnologias Assistivas para Surdos*) project[1] developed at the University of Campinas in collaboration with other institutes. It was created a parallel Portuguese-Libras corpus based on the content of a basic science education textbook, built to be used as a support for Portuguese-Libras automatic translators (MARTINO *et al.*, 2016; KUMADA *et al.*, 2016; MARTINO *et al.*, 2017). During such research, the language experts associated with the project discovered several facial expressions, yet, were never published. The work made in Paiva *et al.* (2018) brings some of those expressions with a focus on intensity (superlative) experimental study.

Based on reference indications, we started our analysis with often-used literature in Libras, the dictionary by Capovilla *et al.* (2017). Composed of a detailed description of Libras grammar and structures, a complete definition of signs, their writing in SignWritting, and translation to Portuguese and English; the dictionary documents more than thirteen thousand signs with individual entries and an analysis of the language, including a list of possible non-manual markers.

[1] A multidisciplinary research group composed of deaf individuals, sign interpreters, linguists, engineers, and computer scientists, that aims to advance the development of assistive technologies for the Deaf. For additional references: http://www.tas.fee.unicamp.br

Another book indication was Quadros e Karnopp (2009). The book is considered a reference for Libras grammar by describing this linguistic system in phonology, morphology, and syntax. Furthermore, non-manual markers are presented in a few images, and their description is detailed in some explanatory speeches.

Even though these studies brought different aspects of non-manual markers to our view, we still felt the need for a more comprehensive survey, that is, a search for more current materials. So, it was made a bibliographic survey of scientific works that had as the topic the facial expressions in Libras, and also has the objective of deepening studying them. As our protocol, we had select two repositories for query (i.e., Google Scholar and SciELO), with the crossing of the following descriptors: non-manual expressions, grammatical facial expression, Libras, Brazilian Sign Language. In Appendix A, we describe the steps we have taken to encounter the works summarized below.

Freitas *et al.* (2014) focus on facial expressions that have morphological and syntactic functions in the Libras' discourse. Based on the works of Quadros e Karnopp (2009), Ferreira-Brito (1990), and Arrotéia (2005), they described facial expression present in association with the syntactic functions of Libras. Such a study was made as part of a proposition of a recognition method for facial expression in Libras.

In (ARAUJO, 2013) was treated the iconicity that exists in sign languages beyond the hands, with a focus on non-manual markings. The literature review compares facial expressions described in works of Baker-Shenk (1983), and Ferreira-Brito (1995), resulting in an extensive list of non-manual markers encountered in Libras. Later in the text, they build a video base to obtain proof that these expressions are not mimic but are part of the Libras' grammar.

Felipe (2013) work on the description of the Libras non-manual markers, presented them as verb-visual units of utterances. After tracing a historical background of non-manual markers in sign language and oral language, it was concluded that in every statement, there is prosody on the corporal and facial, which is realized through non-manual markers. Also, it was affirmed that such expressions are used by deaf and listeners speakers, being a cognitive-discursive strategy that could be considered in a linguistic discourse.

Based on previous works, Xavier (2017) developed a study to compare the basal and the intensified forms of 27 signs of Libras (XAVIER, 2013; XAVIER; BARBOSA, 2013; XAVIER, 2014; XAVIER; BARBOSA, 2014). In their analysis, it was concluded that changes in facial and body expressions were employed consistently throughout all the signs of the experiment when it was produced the intensified form of the isolated signs. Although they describe non-manual articulators (head, eyes, mouth, etc.), their

characteristic aspects were not discussed in particular. Later on, Xavier (2019) analyzes lexical NME of Libras using 368 signals from Xavier (2006) database. The results of this work suggest interesting characteristics of linguistic NMMs, although not all NMMs analyzed are, in fact, lexical. In annexes one and two from this work, a list of NMM encountered in their video set is produced.

The work accomplished in Rezende (2016) is a particular feature of the study of Rezende *et al.* (2016) and Almeida (2014). The work of Rezende (2016) had the objective of translating a video entry of Libras sign using only facial expressions recognition. However, it was not treated facial expressions as non-manual markers or presented a study of facial expressions in Libras. Instead, facial expressions were recognized associated with sign translation. In the same line, Almeida (2014) shows as its objective the recognition of phonological parameters of dynamic signs in Libras. Besides covering manual signs in their Libras literature review part, further on, it was made a description of the phonological nature of non-manual articulators by facial expressions.

Arrotéia (2005) work brings empirical arguments to discuss the possibility of a negative agreement in a sign language by the presence of headshake and negative expression[2]. By comparison to other sign languages, they bring examples of negative non-manual markers in Libras and analyzes the presence of headshake and negative expression. With an extensive dismemberment of non-manual marking in negative sentences, it had concluded the importance of grammatical non-manual marking and their negative interpretation for Libras.

The work Pêgo (2013) deals with mouth-morphemes through the analysis that mouth morphemes have morpho-lexical properties. In that work, it was observed that morpheme-mouth and lexeme-mouth are governed by specific linguistic rules, have coordinated time, and assign particular meanings. It has been noted that some morphemes occur only with mouth movements, others involve different non-manual articulators, such as the head and shoulders. Furthermore, mouth morpheme that simulates a breath, but with tight lips, takes on inflated cheeks or decreased air in the cheeks indicates the amount associated with specific hand signs can change the meaning. Thus, the different possibilities of using morpheme-mouth and some other few combinations of facial expressions were described. Therefore, both works Arrotéia (2005) and Pêgo (2013) deal with non-manual markings; however, only specific articulators heavily combined with manual signs. Also, they do not describe the full range of compound face expression cases. Such works were not added to our analysis of the facial expressions in section 3.3.

It is noted here the diversity of approaches in the study of facial expressions,

[2] In Arrotéia (2005), it is defined as negative expression the prototypical face of frown and crocked mouth down or rounding of the lips, and a slight lowering of the head.

terminologies, and applications. Thus, we feel justified the need for a study of broad non-manual markings of Libras and for standardization, which in the future can be used as labels for our FER system.

3.2 Grammatical and Affective Facial Expressions in Libras

Studies at the phonological level of sign language began in the 1960s, with the pioneer William Stokoe Jr. (STOKOE, 1960) by describing what is considered three primary parameters (movement, location, and hand configuration). In the 1970s, the orientation of the palms of the hand and non-manual expressions are indicated as secondary parameters integrating these phonological studies (BATTISON, 1974; BAKER; PADDEN, 1978; LIDDELL, 1978).

In this evolution of studies on sign language, non-manual expressions are considered minimal units of sign languages, analogous to phonemes, although some authors prefer to use concepts such as quirema, visema or others to mark the specificity of sign languages and to dissociate the idea of the phoneme that is traditionally linked to sound (for more information on this terminological conflict, we suggest reading Capovilla (2013)).

Non-manual markers integrate the movements of the face, head, and trunk. However, it is currently demonstrated that the signing extends to the entire person's body, for example, when the sign interpreter lends its body to the character during phenomena known as classifiers or partitioning of the subject's body (BERNARDINO, 2012; CARNEIRO, 2015; ALBRES *et al.*, 2015). In this research, we are interested in analyzing the facial expressions of Libras, mainly. Facial expressions may change widely in Libras (Figure 3.1) and due to their relevance in the sign language phonology, non-manual markers have a distinctive function, that is, an unique facial expression may be combined to many different manual signs and the meaning of the same manual sign can vary significantly. In line with the literature review made by Souza (2014) on NMM of Libras, our studies unveiled that there is still limited documentation regarding the list of facial expressions that carry relevant meaning in Libras and how they combine with manual signs.

In this direction, it is possible to find manual, non-manual, and multichannel/multimodal signs in Libras. The manual signs would be those performed with the use of hands unaccompanied by a non-manual expression (or associated with a neutral expression). Non-manual signs are understood as those in which only non-manual markers are presented, such as in the signs of "Steal" and "Sex" in Libras (SANTOS; XAVIER, 2019) or pronouns performed merely with the eye (FELIPE, 1997). There are also multichannel/multimodal signs; that is, they bring together the use of hands, facial, and body

Figure 3.1 – Image (A) shows the sign "Why" in an interrogative sentence. In the image (B) is performed the sign "Which" in an interrogative sentence. In the image (C), it can be observed the sign "Candy", in an affirmative sentence. In the image (D), there is the sign "Thinking" being presented in a doubt sentence. Source: Corpus of the research itself (SILVA *et al.*, 2020b).

expressions. Our studies unveiled that there is still limited documentation regarding the list of facial expressions that carry relevant meaning in Libras and how they combine with manual signs.

In Libras, facial expressions that convey an idea of feeling and emotion, are called Affective Facial Expressions (AFE). Affective facial expressions can start before a specific sign and end after the sentence has been completed. In other words, AFEs modulate the whole sentence, modifying the full meaning of a sequence of signs (ARAUJO, 2013). AFEs are adopted, for example, when the signer communicates ideas sarcastically or when he/she is describing a sad event(PIMENTA; QUADROS, 2008; SOUZA, 2014; SANTOS; XAVIER, 2019). A visual characteristic of AFEs is that they employ an integrated set of facial muscles.

Grammatical facial expressions (GFE) (also referred to as non-manual markers (SOUZA, 2014) or linguistic expressions to identify their grammatical function (PIZZIO *et al.*, 2009)) in Libras are expressions that typically occur at specific times in a sentence, or they are associated with a particular execution of sign (SOUZA, 2014; FREITAS *et al.*, 2014; FREITAS, 2011). According to Quadros and Karnopp (2009, 2014), and Arrotéia (2005), GFEs are defined in the structure of Libras, at the morphological level [3] and the level of syntax[4], at the phonological level according to Quadros and Karnopp (2014), and other authors who indicate that non-manual markers are considered a phoneme, and are obligatory in certain contexts. Another possible classification of facial expressions would

[3] Morphology is the study of words and their relationship to other words in the same language. The morphological level of a language is the level at which meaning can be assigned to parts of words and the level that describes how morphemes (the smallest meaning elements of words) are combined to make a word (COOK; THOMAS, 2005).

[4] The term syntax refers to a set of rules, principles, and processes that govern the structure of sentences in a given language. The syntactic level of structure concerns the structure of the sentence, i.e., the categories of words and the order in which they are assembled to form a grammatical sentence (COOK; THOMAS, 2005).

be associated with the lexical and syntactic level (SANTOS; XAVIER, 2019). So they can be used to change the meaning of a sign or can transform the sentence of the phrase.

At the morphological level, non-manual markers may usually accompany adjectives and nouns. Yet, we can find facial expressions in some verbs, as well as in classifiers associated with the action, e.g.: "Run"; "Contaminate"; "Exist", etc. When the NMMs follows an adjective, it can determine the degree of intensity. Whereas when the non-manuals accompany a noun, this may indicates the degree of size. Thus, with the presence of NMM, can have a superlative and comparative construction of superiority and inferiority (QUADROS; KARNOPP, 2009). Still at the lexical level, it is important to highlight the presence of lexical mouth components that may be associated with a movement of the mouth during the performance of the sign, including here the linguistic loans of the oral language recognized in the attempt to oralize or articulate the mouth with the figure of the word of the oral language (SANTOS; XAVIER, 2019). Furthermore, at the level of syntax, non-manual markers are responsible for constructing sentences of an affirmative, negative, interrogative, relative, conditional, topic, and focus types (QUADROS; KARNOPP, 2009).

We separated the facial articulators in classes to clearly define their role in the discourse. It is possible to found names for the purpose that facial expressions assume in sign language, as morphemes (ARAUJO, 2013), syntactic (QUADROS; KARNOPP, 2009; PIMENTA; QUADROS, 2008), phonetic-phonological, lexical (XAVIER, 2017; XAVIER, 2019), or precise terms formed by the combination of Greek and Latin morphemes, for example, *MascarEmas-PersonalÍculos*[5], and/or *MascarasCinesEmas-PersonaMotusÍculos*[6] (CAPOVILLA; GARCIA, 2011; CAPOVILLA; GARCIA, 2012; DOMINGUES, 2015). Notice that many of these terminologies bring conflicts and different definitions depending on the reference adopted. So our class types and nomenclature were based on the literature. However, We used the logic of sets to observed intersections between the facial articulators' roles. For example, in the lexical realm, FE is used to increase the adjective intensity and differentiate the sign's meaning, which has completely different roles and articulators for the same class name. That can confuse an automatic recognizer of facial expressions in sign language. Thus, we divided the class of FE that have a lexical role into new facial expressions grammatical sets, as discussed below.

Observing the different properties of grammatical facial expressions we chose to categorize them into *Grammatical facial Expression for Sentence* (GES), *Grammati-*

[5] From Greek morphemes: *Mascar* (mask), *Ema* (minimum unit), and; from Latin morphemes: *Personal* (person), *Ículos* (minimum unit) (DOMINGUES, 2015).

[6] From Greek morphemes: *Mascaras, Mascas* (mask, masks), *Cines* (movement), *Emas* (minimum unit), and; from Latin morphemes: *Persona* (personal, person), *Motus* (movement, motion), *Ículos* (minimum unit) (DOMINGUES, 2015).

cal facial Expressions of Intensity (GEI), *Grammatical facial Expressions of Homonymy* (GEH) and *Grammatical Facial expression of Norm* (GEN).

Grammatical facial expression for sentence, or GES, defines the type of sentence that is being signed. Also, GES are the most mentioned in the literature (ARROTÉIA, 2005; PIMENTA; QUADROS, 2006; QUADROS; KARNOPP, 2009; ARAUJO, 2013; FREITAS *et al.*, 2014). In Libras, according to Quadros and Karnopp (2009), and Ferreira-Brito (1995), there are GES markers that are expressed by the face and head movements:

- **WH-question** (WH): generally used to represent interrogative pronouns such as *what*, *who*, *when*, *why*, *where* and *how*;
- **Yes/No question** (YN): used when the question being asked has an *yes*/*no* answer;
- **Doubt question** (DQ): it is used to emphasize the information that will be supplied;
- **Topic** (T): when one of the sentence's constituents is displaced to the beginning of the sentence;
- **Negative** (N): used in negative sentences;
- **Assertive** (A): used when making statements;
- **Conditional clause** (CC): used in subordinate sentence to indicate a prerequisite to the main sentence;
- **Focus** (F): used to highlight new information into the speech pattern;
- **Relative clause** (RC): used to provide more information about something.

Thus, GES mark the modalities of statements. For example, Figure 3.2 presents the Libras' sign "not know". We observe a negative non-manual feature marked by the movement of the head sideways, the frowning of the eyebrows and the slightly curved down lips.

In its turn, lexical grammatical expressions are linked, as the name suggests, to the lexicon (PIMENTA; QUADROS, 2006). Our survey showed that a good portion of the studies (PAIVA *et al.*, 2018; XAVIER, 2017; PÊGO, 2013) restrict the analysis and description of lexical expressions to intensity expressions, or Grammatical Facial Expressions of Intensity (GEI).

Figure 3.2 – Performance of the signs in Libras using a negative GES. Source: Corpus of the research itself.

GEI differentiate the meaning of the sign assuming the role of quantifier. They constitute lexical components that mark dimension on adverbs or adjectives (QUADROS; KARNOPP, 2009). For example, when analyzing Figure 3.3, it is possible to observe that the same sign associated with the word "beautiful", in Figure 3.3 (B), can have its meaning reduced to "cute", in Figure 3.3 (A). The main changes can be observed in the eyes, the eyebrows, and the mouth. In Figure 3.3 (C), the sign "pretty" is represented by the evident changes in the raised eyebrows and open mouth.

We highlight that, in the analysis of the literature, we found two other subcategories of facial expressions associated with lexical and grammatical expressions. However,

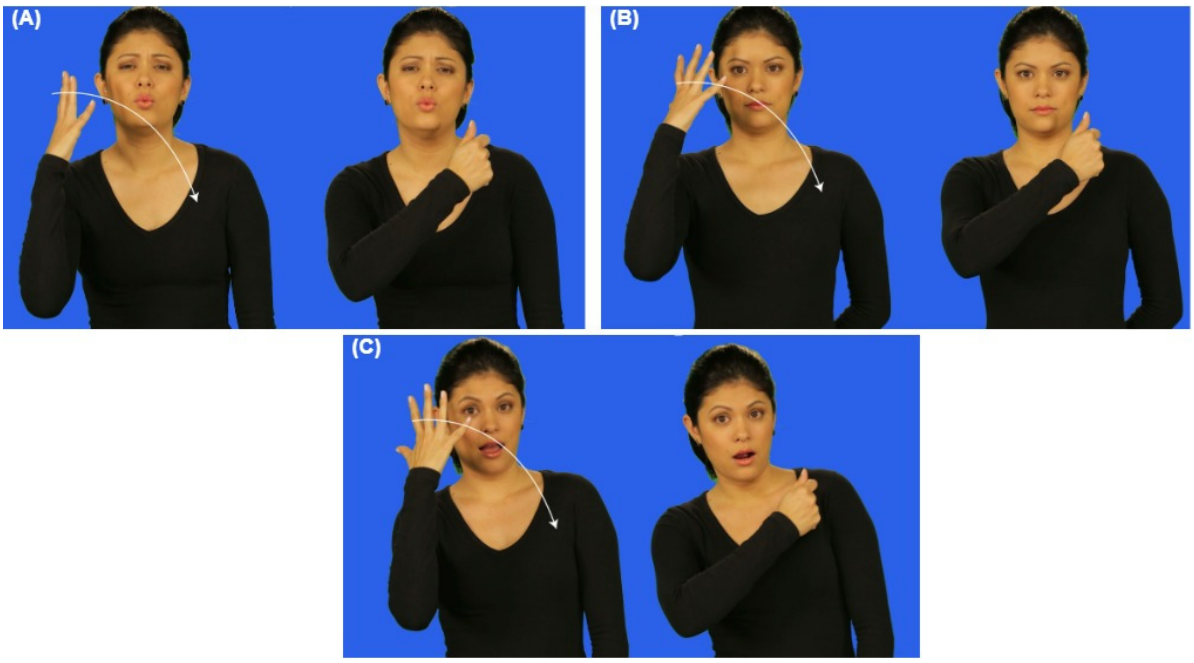


Figure 3.3 – Performance of the signs in Libras using GEI. Source: Kumada *et al.* (2016), Kumada (2016).

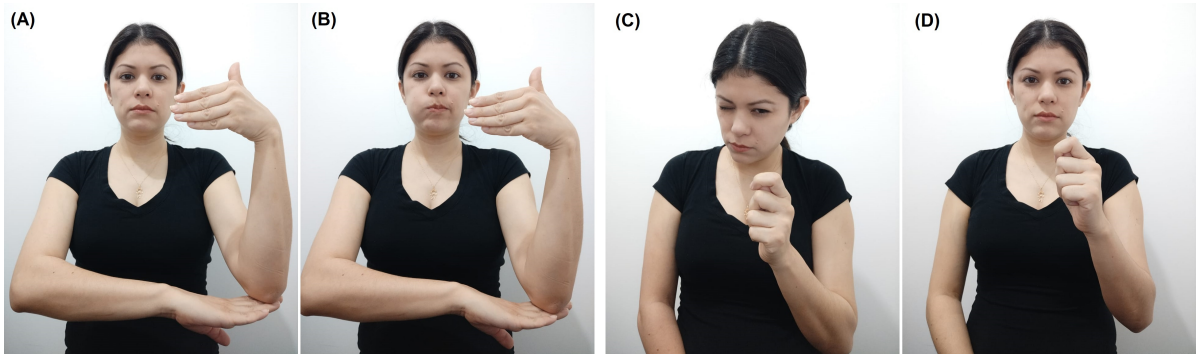


Figure 3.4 – Performance of the signs in Libras using GEH and GEN, respectively. Source: Corpus of the research itself.

we did not found any proper classification for them, and we named them as the Grammatical facial Expressions of Homonymy (GEH) and Grammatical facial Expressions of Norm (GEN).

It is known that facial expressions as one of the parameters of Libras function as distinctive features, determining lexical contrastivity, for example, of concepts such as "occupied" and "not-can" in Libras, altered merely by the non-manual marker (XAVIER; BARBOSA, 2014). Under this perception, the GEH differentiate signs with the same manual sign. They help to define the meaning of a sign and they are not associated with the change of sentence or intensity in their qualification. For example, Figures 3.4 (A) and (B) show signs that have the same manual sign but have different facial expressions. In (A) the interpreter is signing "Hotel", while in (B), the sign "motel" is shown[7]. Their distinction is only based on one of the cheeks and mouth action. In this category, we would also include the lexical oral components that, by employing the reference to the figure of the spoken word, can act in the distinction from one sign to another. According to Pimenta e Quadros (2008), polysemy (same words with many meanings and the same origin) or homonymy (words with the same or similar phonetic identities, but with different backgrounds and meanings) are also found in Libras; still, studies about their respect are scarce. So we denote them GEH, because homonymous terms (especially homonymous homographs)[8] can be associated with subtle changes in the pronunciation of a word, such as the occurrence of these signs in Libras[9].

Finally, we define the Grammatical facial Expressions of Norm (GEN) as facial

[7] For authors like Santos e Xavier (2019), the "motel" sign would be associated with the simultaneity of the non-manual sign of "sex" with the manual sign of "hotel"; however, there would not classify facial expressions of signs in this conditions

[8] Due to the difficulty of retrieving records of the origins of Libras' signs to identify whether they could be categorized as homonymous or polysemic, as well as considering the absence of a consolidated Libras written record, it becomes complex to specify or differentiate in this research the polysemic signs, homographs or homophones homonyms.

[9] It is noteworthy that, in addition to facial expressions, we do not ignore that in sentences with the presence of a homonym or polysemy, the context can also be decisive for identification.

expressions that are part of the signal by definition, i.e. by norm. Thus, when a GEN sign is performed without the facial expression that defines it, the sign loses its meaning. For example, Figure 3.4 (C) presents the Libras sign for "magnifying glass". If the facial expression is not produced, as shown in Figure 3.4 (D), the gesture has no meaning in Libras. Here, we highlight that the articulator that functions as a distinctive feature is one closed eye, not a buccal lexical component, as studies on the upper part of the face in Libras have been more frequently associated with GES (XAVIER, 2019) with little discussion of their influence on the lexical level.

According to Felipe (2013), many of the non-manual articulators found in Libras are also present in other sign languages. For example, *yes/no-questions* in American SL (ASL) are associated with raised eyebrows, head tilted forward and widely-opened eyes. Similarly, *wh-questions* are represented by furrowed eyebrows and head forward. Topics are described by raised eyebrows and head slightly back, and negations are expressed with a head shake. The head pose and eye gaze describe dialog during a story narration (ANTONAKOS *et al.*, 2015; LIDDELL, 1980). In the German SL (DGS), a change of head pose combined with the lifting of the eyebrows, corresponds to a subjunctive. Lip pattern, tongue, and cheeks that are not related to the articulation of words can provide information redundant to gesturing to support differentiation of similar signs, i.e., solve the ambiguity (AGRIS *et al.*, 2008). These facial expressions function intersections reinforce the hypothesis that is possible to expand a classification model in Libras to other sign languages.

3.3 Proposed Taxonomy for Facial Expressions in Libras

As a result of our survey, Table 3.1 compile and standardizes our findings on the mapping of Libras facial expressions considering the references which we found to be the most relevant and complete in the context of our study: Quadros e Karnopp (2009), Pimenta e Quadros (2006), Araujo (2013), Freitas *et al.* (2014), Capovilla *et al.* (2008), Capovilla *et al.* (2017), Kumada *et al.* (2016), Paiva *et al.* (2018), and Xavier (2019). Aside from showing a standard vocabulary for facial expressions, we also associate the facial articulators described by each work. When analyzing the number of facial expressions, by comparison, there are many differences between the number of facial articulators described by each author.

Table 3.1 separate the non-manual markers in head and face, the latter being further split into upper and lower face parts. That was made to organize the articulators. Yet, despite the more significant number of lower articulators, it is noted that expressions associated with syntax use higher articulators more (CRASBORN, 2006).

Table 3.1 – Facial Expressions in Libras as described in the literature

Facial Expressions	Authors					
Face						
Upper face						
Frown	Q	A	T	F	C	X
Raised eyebrows	Q	A	T	F	C	X
Joined eyebrows			T		C	
Left / Right eyebrow raised			T			X
Wide open eyes	Q	A	T	F	C	X
Slightly closed eyes	Q		T	F	C	X
Closed eyes		A			C	X
Open eyes		A	T			X
Left / Right eye closed			T			
Look at the speaker		A				
Direct the eyes	Q		T			X
Nose wrinkle	Q		T			
Lower face						
Inflated cheeks	Q	A	T		C	X
Contracted cheeks	Q	A	T		C	X
Contracted cheeks and projected lips	Q	A				
Contracted lips			T	F		X
Projected lips			T	F	C	
Only left / right cheek inflated	Q		T			X
Run the tongue against the lower part of the cheek	Q		T			X
Smile with apparent teeth			T		C	
Crooked mouth up		A	T	F	C	X
Crooked mouth up laterally			T			
Crooked mouth down		A	T	F	C	X
Crooked mouth down laterally			T			
Contraction of the upper lip	Q		T			
Lower lip pressed by upper teeth			T			X
Semi-open mouth (blowing)			T			X
Sibilant tongue[10]		A	T			X
Lips apart			T		C	X
Open mouth		A	T	F	C	X
Closed mouth			T			X
Clenched teeth			T		C	X
Swinging alveolar tongue[11]			T			X
Tongue in lisp position			T		C	X
Mouth movement (mouthings)		A				
Chewing movement		A				X
Snap of the lips						X
Head						
Balance back and forth (yes)	Q	A	T	F		X
Quick nod			T		C	X
Balancing sideways (no)	Q	A	T	F	C	X
Brief and upward movement of the head		A	T		C	X
Forward lean	Q		T	F	C	X
Tilt to the side	Q	A	T			X
Tilt back	Q		T	F		

Q - Quadros e Karnopp (2009); A - Araujo (2013); T - TAS (2012), Kumada *et al.* (2016), Martino *et al.* (2017), Paiva *et al.* (2018); F - Freitas *et al.* (2014); C - Capovilla *et al.* (2008), Capovilla *et al.* (2017); X - Xavier (2019).

10-In other words, open mouth and tongue in movement.

11-This term refers to the region within the mouth in the inside margin of the upper central incisors.

In Libras dictionaries, in particular the dictionaries authored by Capovilla and colleagues (CAPOVILLA *et al.*, 2008; CAPOVILLA *et al.*, 2017), we observe that the description and the quantity of facial expressions vary according to the dictionary edition. For example, in Capovilla *et al.* (2017) the facial expressions are described by what part of the face moves and the possibilities of movement. In contrast, in a previous edition of the dictionary (CAPOVILLA *et al.*, 2008), only bring texts like "sad expression" or "interrogative expression", as if there were only one possible facial expression form that accompanies a sign. In spite of the improvement in the latest edition, the facial expression descriptions continue to be vague. Analogously to Xavier's (2019) criticism is that the explanation of the signs is not always accompanied by a mention of the articulators involved. As in the case, for example, of "Headache" where the only description is "facial expression of pain", which is not indicative of what specifically articulators would be involved. This can be a kind of evidence that even with the evolution of language over time, there is still uncertainty about the best way to standardize the non-manual articulators of the language.

Table 3.1 highlights that different authors refer to distinct sets of relevant facial expressions and there are some expressions that are not mentioned by one or more authors. For example, the "crooked mouth down laterally" expression was considered relevant by the TAS Project group (KUMADA *et al.*, 2016; MARTINO *et al.*, 2017; PAIVA *et al.*, 2018), but it is not explicitly mentioned by other references. In addition, the authors do not have a consensus in recording the description of the relevant NMM in Libras. Table 3.1 presents a standardized nomenclature that takes into account the movement performed by the face and the head.

A key aspect in the development of our sign-language FER model is to identify the group of facial expressions and head movements that can be combined to convey meaning in Libras. As a result of our survey, we were able to associate combinations of the primitive non-manual markers listed in Table 3.1, to specific semantic functions, as shown in Table 3.2. The making of such classification is to describe the facial articulators pertinent to each defined class. In a way, the separation by class could auxiliary the modelling of language behaviour. As discussed in Section 3.2, Table 3.2 includes two grammatical categories that were not previously named in the literature, the Grammatical facial Expression of Homonomy (GEH) and the Grammatical facial Expression of Norm (GEN).

Table 3.2 was constructed from the analysis of works Pimenta e Quadros (2006), Capovilla *et al.* (2008), Freitas *et al.* (2014), Paiva *et al.* (2018).

Table 3.2 – Taxonomy of Libras' facial expressions accordingly with the non-manual markers classifications.

AFE	Left / Right eyebrow raised Raised eyebrows and wide open eyes Raised eyebrows, wide open eyes and open mouth Slightly closed eyes and crooked mouth up Smile with apparent teeth Smile with apparent teeth and open mouth Lowered eyebrows and crooked mouth down Frown and contraction of the upper lip Crooked mouth up laterally		
GFE	GES	WH	Brief and upward movement of the head and frown Tilt back,frown and projected lips
		YN	Brief and upward movement of the head and raised eyebrows Tilt to the side, frown and projected lips Balancing sideways, frown and projected lips
		DQ	Frown, slightly closed eyes and contracted lips
		T	Brief upward and forward movement of the head, raised eyebrows, open mouth, projected lips Quick nod, brief upward movement and wide open eyes Quick nod, brief upward movement, raised eyebrows and wide open eyes Quick nod, brief upward movement, raised eyebrows, open mouth and projected lips
		N	Crooked mouth down; Quick nod, frown and crooked mouth down; Head balancing sideways.
		A	Balance back and forth of the head
		CC	Brief and upward movement of the head and raised eyebrows
		F	Brief upward and forward movement of the head, raised eyebrows, open mouth, projected lips Quick nod, brief upward movement and wide open eyes Quick nod, brief upward movement, raised eyebrows and wide open eyes Quick nod, brief upward movement, raised eyebrows, open mouth and projected lips
		RC	Raised eyebrows
	GEI	Frown < Frown and Slightly closed eyes Inflated cheeks and semi-open mouth < inflated cheeks, semi-open mouth and frown Contracted cheeks and frown < contracted cheeks Contracted lips and frown < contracted lips Projected lips and frown < projected lips; Open mouth and frown < open mouth Crooked mouth up< Smile with apparent teeth Quick nod < balance back and forth of the head	
	GEH	Inflated cheeks Only inflated right cheek Open mouth Mouth crooked up Lips projected	
	GEN	Closed left eye, eyebrows raised, brows down, lips contracted, lips projected, mouth crooked up, mouth crooked down, teeth apparent, mouth open and teeth apparent, cheek inflated, cheek contracted, head up and down, head movement to turn from side to side.	

Semantic Functions: AFE- Affective Facial Expression; GFE- Grammatical Facial Expression; WH- WH-question; YN-Yes/No questions, DQ- Doubt question, T- Topic, N- Negative, A- Assertive, CC- Conditional Clause, F- Focus, RC- Relative Clause.

3.4 Concluding Remarks

This chapter presented the results of a concise linguistics literature survey that aimed to identify what are and how facial expressions are characterized in Libras. The results presented in this chapter represent the theoretical basis to build our sign-language FER model.

The survey showed that there is no universal standard regarding the terminologies adopted to describe facial expressions in Libras, what generates a challenging aspect of designing a computational application for recognition of Libras' non-manual markers objectively.

As a contribution of the present work we interpreted the multiple terminologies and we derived a comparison among different authors. From the literature, we identified the typical semantic functions of non-manual features in Libras, but our study also unveiled new categories that were not previously formally described in the literature. We presented a taxonomy of semantic functions in Libras and their corresponding non-manual markers.

Alongside to defining terminologies, is the necessity of a transcription tool with a structure and capability of associating visual features in a machine-readable element of Libras' facial expression. In the next chapter, we will cover the whole transcription scheme in detail.

4 Annotation Models for Libras' Facial Expressions

Our design of a sign language FER model was intertwined with its envisioned applications. We considered not only that the output of our FER model could be processed by other computational systems (machine-readable requirement) but also that its output could be processed by researchers that study the language (human-readable requirement).

In this chapter, we present the main aspects that guided our design through a review of the annotation schemes and the computational tools that are typically adopted by the researchers in Linguistics, with a focus on Libras (Sections 4.1 to 4.3). Then, we propose an annotation model for facial expression in Libras based on FACS (Section 4.4). Finally, in Section 4.5, we present a framework for automatic transcription of facial expression in Libras that can be adapted to any sign language.

4.1 Sign Language Transcription

Transcription is known as the process of representing something in a written form, like language, audio, images, genetics, among others (CROWTHER, 1995). Automatic image transcription systems are typically based on computer vision techniques that extract relevant features from images, associated with supervised machine learning algorithms that translate the features into tags, learned from previous examples (LI; WANG, 2003; MURTHY *et al.*, 2015). The automatic transcribing process of an image is also known as automatic image annotation, automatic image tagging or linguistic indexing.

Transcription and linguistic analysis have a close relationship due to the transcription aptitude for a liable and faithful record of the data, also, the capacity to register what is significant and not register absolutely everything. Automatic speech transcription is a widely discussed topic (XIONG *et al.*, 2018; ZEYER *et al.*, 2017; WANG; BROWN, 2006; WOODLAND; POVEY, 2002; SEIDE *et al.*, 2011; SAINDON *et al.*, 2004), whilst automatic transcription of sign language is still relatively little explored (SHI; LIVESCU, 2017; AHMED *et al.*, 2017; NEIDLE *et al.*, 2018; RYBACH *et al.*, 2006). The first is facilitated by the availability and adaptation of alphabetic systems that support spoken languages. The second, however, is a more complex problem, since the existing writing systems for sign language are not trivial. The gloss system[1] is accepted with reservations

[1] In our study, we are calling the linguistic notation system used in Brazil to register Libras through written Portuguese (using capital letters) a gloss. We prefer the term gloss over word since it is

by the linguistics community. Mainly, systems that automatically transcribe non-manual markers from a video entry are unheard of.

4.2 Libras' Facial Expressions Annotation Models

While the registration and transcription of samples of the spoken language are easily obtained through writing using the alphabetic system of the language itself, the registration of a sign language is only possible through technological resources such as video recording and its transcription, which still does not have consensus even among scholars in the field, being carried out through writing systems designed specifically for this purpose (such as the Sign Language Writing and the Hamburg Notation System), or via a word notation system. According to Bois (1992), within linguistics, coding involves registry in higher levels of interpretation and analysis, while transcription is directly observed.

The Sign Language Writing (ELiS), a sign language writing system, was created by Barros *et al.* (2008), Barros (2016) based on the use of 95 visogrammes and adapted to a TrueType font, allowing its writing from a standard computer keyboard. According to the author, ELiS is organized according to the six types of signs: monomanual, symmetrical bimanual, asymmetric bimanual, almost symmetrical bimanual, with support and compound hand. Nevertheless, it still appears to have limitations for inserting non-manual marks.

Additionally, the Hamburg Notation System for Sign Languages (HamNoSys) (HANKE, 2004) is a system to describe signs on a mostly phonetic level. It was designed to fit research where transcriptions are exact. The symbology for a hand form resembles remarkably the actual handshape desired. HamNoSys provides extended symbols for manual parameters of sign language, and despite being possible to note down facial expressions, such a feature is still being improved. In Table 4.1 is presented the symbols for non-manual markers in HamNoSys.

There is also Sutton SignWriting (or just SignWriting)(SUTTON, 1995), wherein Brazil is known as "Escrita em Sinais" (in Portuguese), it consists of a visual representation of the signs, through a schematic graphic process capable of describing, through

known that when dealing with two languages, it is possible that in one of these, there is an expression represented by a word, and in another, the concept generates more than one word. For example, in English, the word "Sunglasses" is represented by more than one Portuguese word "Óculos de sol". In Libras, this also happens, and the reference the gloss is thus made more suitable for such correspondence. In addition, the adoption of gloss in sign language allows adding information to this annotation, such as indicating the interlocutor, for example, in the register of 1sPERGUNTAR2s (FELIPE, 1997) it can be learned that the first singular person performs the action of the verb ASK for the second person singular. In short, the concept of disallowance becomes more comprehensive than that of the word.

Table 4.1 – Non-manual markers representation in HamNoSys

Symbol	Description
	Head
	Forehead
	Under the nose
	Mouth
	Chin
	Cheeks
	Shoulders
	Middle torso
	Lower torso

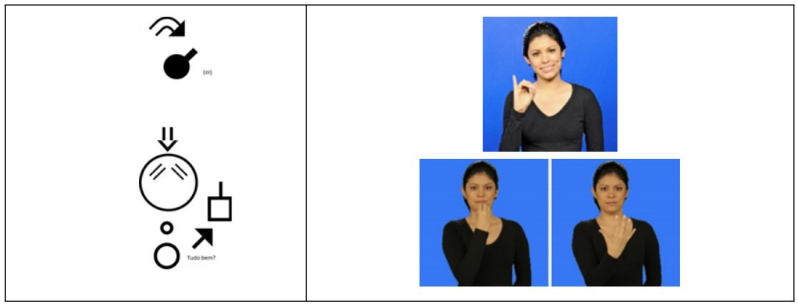


Figure 4.1 – Representation in signwriting and Libras signs for the phrase "Hi, How are you?". Source: SignWriting images elaborated by Maria Salomé Soares Dallan (2017) and Libras images extracted from Kumada (2016).

the written form, the fundamental gestural units, their articulators and relationships. An example of SignWriting is illustrated in the observation of Figure 4.1.

Even though these annotation models are applied in Libras, it is expected that researchers opt for the "word notation system" (FELIPE, 1997), which uses the words of the oral language, in Libras context Portuguese, to register the signs. The annotation of non-manuals markers in Libras is typically performed through codes and non-standardized orthographic systems.

Ferreira-Brito (1995) and Souza (1998) were the first to propose transcription tags for non-manual markers. Focusing on the function that an expression exhibit in a sentence, they adopt a line above the gloss to indicate the type of non-manual that was observed. The labels used are "_?" for a question, "_!" for exclamative, "_ñ" for negation,

"_t" for a topic, "_EFp" for illocutionary force[2], and "_EFo" for the case of an order[3]. For instance, the phrase:

I will buy a pen today. (4.1)

when annotated using the system proposed by Ferreira-Brito (1995), becomes:

$$\overline{\text{PEN}}^{\text{t}}, \quad \text{I} \quad \text{BUY} \quad \text{TODAY}.$$

Considering the linguistic literature on American Sign Language (ASL), Quadros and Karnopp (2009) proposed representing facial expressions and head movements, including the grammatical aspect involved. The proposed system delimiters the glosses using the smaller-than (<) and greater-than (>) signs and subscribe them with a code to represent a non-manual marker category (see Table 4.2). Considering the previous example, the annotation process, according to Quadros and Karnopp (2009) results in:

$$< \text{PEN} >_{t}, \quad \text{I} \quad \text{BUY} \quad \text{TODAY}.$$

Both of the annotation systems presented adopted an elaborate coding scheme focusing on analysis by humans, sign language students, and linguistic researchers. That makes them too complex to be parsed into a machine-readable code. Besides, by building upon the grammatical function of non-manual markers, the other roles (as mentioned before, as of intensity or emotional state) of facial expressions may not be contemplated in the annotation model.

Table 4.2 – Transcription scheme proposed by Quadros and Karnopp, 2009

Grammatical Function	**Description**	**Coding**
Question	Record of question	$<>_{qu}$
	Record of a yes or no type of question	$<>_{sn}$
Assertion	Eyegaze in a grammatical agreement	$<>_{do}$
Negation	Head shake	$<>_{n}$
Topic	Tagging topic	$<>_{t}$
Foco	Head Movement	$<>_{mc}$

[2] Illocutionary force refers to a speaker's intention in delivering an utterance or to the kind of illocutionary act the speaker is performing (NORDQUIST, 2018; ALSTON; ALSTON, 2000). For instance, when somebody says, "Is there any napkin?" at the dinner table, the illocutionary act is a request: "please give me a napkin" even though the *locutionary act* (the literal sentence) was to ask a question about the presence of a napkin. The *perlocutionary act* (the actual effect), might be to cause somebody to pass the napkin.

[3] Order in this sense may so be comprised within the Imperative realm, which not only commands or orders but also requests, pleas, appeals, and other linguistic expressions of willing or wishing something to be done or not to be done. The differences between these expressions are not logical, but of a psychological character (JØRGENSEN, 1937).

As a more sophisticated transcription system, McCleary and Viotti (2007) and Mccleary *et al.* (2010) proposed an annotation model that included the description of non-manual markers and facial expressions. Such a system define a description tier for different aspects of the signs, as shown in Table 4.3.

To date, this was the only manual transcription system marking at first non-manual articulators, and secondarily, manual signs, gloss, and translation from Libras to Portuguese. However, the system does not embrace all the non-manual markers we presented in Section 3.3.

Finally, there exist textual transcription systems, in which the non-manual markers are described using templates as "closed expression" or "happy" (CAPOVILLA *et al.*, 2017). However, without objective guidelines, such systems are subject to the different interpretations of the annotators (FRYDRYCH, 2010).

Table 4.3 – Transcription scheme proposed by McCleary e Viotti (2007), Mccleary *et al.* (2010)

Tier	**Description**	**Controlled vocabulary**	**Acronyms**
NMM-Gloss (Non-manual Marker Gloss)	Recording of signs that are performed only by non-manual features		GLOSS
Eyelids	Registry of eyelid configurations and movements	blink	[p]
Eyegaze	Record of settings and eye movements	right	>
		left	<
		up	^
		down	v
Eyebrows	Eyebrow settings log	raised	/\
		frown	\/
		raised eyebrows and frowning	/*\
Head	Record of settings and head movements	forward	F
		right	R
		left	L
		up	U
		down	D
Mouth pictures	Record of visually perceptible mouth movements that are related to Brazilian Portuguese phonemes (visemes)		[phonemes]
Mouth Gestures	Record of oral gestures that have no relation to the Portuguese language	open	OF
		closed	CO
Transcription	Translation to Portuguese		Portuguese

4.3 Video Annotation Tools

Video annotation tools are computational applications to manually or semi-automatically annotate and transcribe video recordings. They typically present a tier-based data model that supports multi-level annotation of time-based media. In sign language research, video annotation tools are widely used to conduct detailed studies about

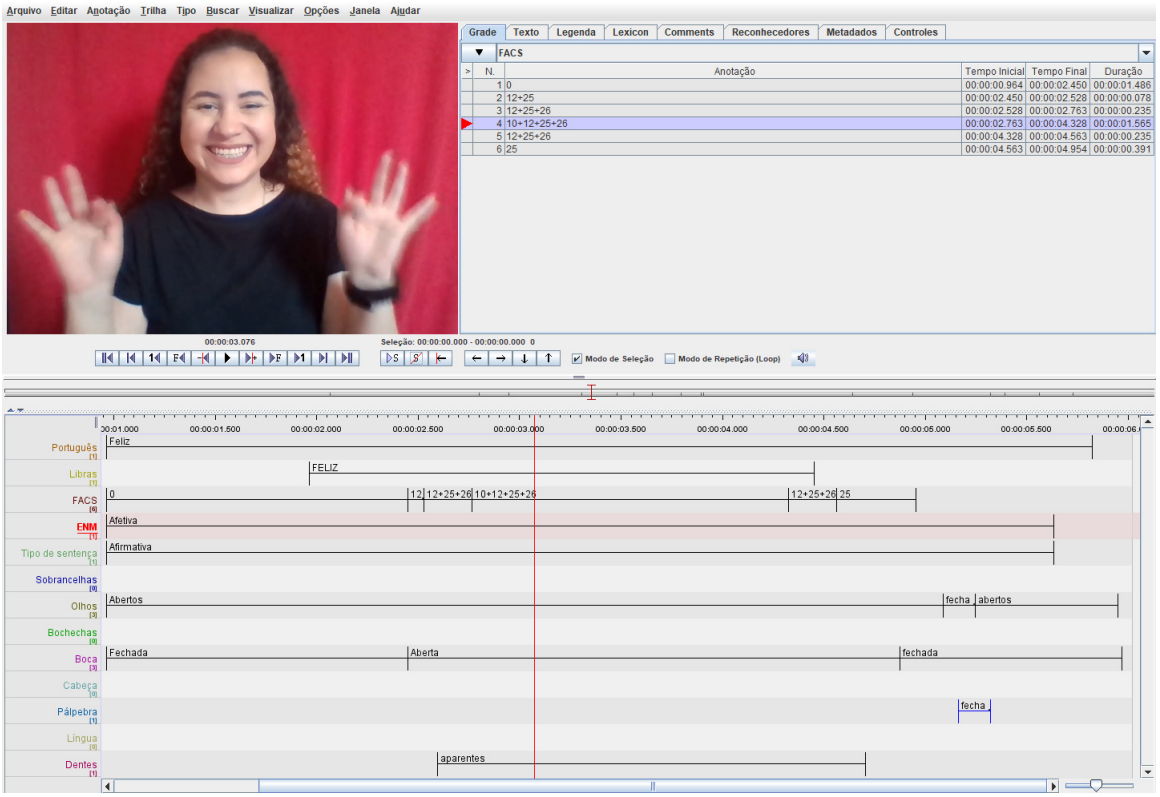


Figure 4.2 – Example of the use of ELAN (Eudico Annotation Tool) with multiple tiers to describe different aspects of a sign. Extracted from the research itself.

the gestures, the signs and their components. Such computational tools typically allow searches and reports of various types. In Figure 4.2 we observe an example of the use of multiple tiers to annotate different aspects and timings of the sign recorded in a video.

Examples of computational annotation tools adopted by sign language researchers include:

- **Annotation of video and language data (ANVIL)** (KIPP, 2001) is a video annotation tool developed for the purpose of annotating non-verbal communication. It allows multilayered annotation based on user-defined coding scheme notes and it is available for educational and research purposes[4].

- **Computerized Language Analysis (CLAN)** (CONTI-RAMSDEN, 1996) program is an editor for creating and analyzing transcripts of CHAT or CA (Conversation Analysis) format. It is distributed upon agreement of exchange information in the database with the project Child Language Data Exchange System (CHILDES)[5].

- **SignStream** (NEIDLE *et al.*, 2001) was designed to facilitate annotations videos from American Sign Language by Boston University. A vast corpus of annotated

[4] http://www.anvil-software.org/
[5] https://childes.talkbank.org/

images in ASL is available in SignStream. It is available for educational and research purposes[6].

- **TRANSANA** (WOODS; FASSNACHT, 2007) is a qualitative data analysis software in which the user can analyze and transcribe multiple camera videos. This product is distributed by Digital Insight Project[7].
- **EUDICO Language Annotator (ELAN)** (BEREZ, 2007) was created for linguistic analysis by the Max Planck Institute for Psycholinguistics. It allows manual annotation in multiple videos, multiple tiers and extract metadata for qualitative and quantitative analysis. It is available as a free and open-source software[8].

Among the presented tools, ELAN is broadly adopted by the scientific community (BEREZ, 2007; LEMOS, 2012; PAIVA *et al.*, 2016), and by Libras linguists (MCCLEARY; VIOTTI, 2007; MARTINO *et al.*, 2017; PAIVA *et al.*, 2018).

4.4 Facial Action Coding Association with Facial Expression in Libras

Even with the advance in computational annotation tools, the transcription systems in Libras did not evolve alongside. The annotation models presented in Section 4.2 could not be easily processed by FER algorithms due to the lack of objective parameters to label the facial expressions, and their annotation complexity. We argue that there is a significant advantage in adopting the Facial Action Coding System (FACS) to describe facial expression in Libras. FACS has the potential of unifying the different terminologies used by each author to describe and map facial markers, facilitating not only the development of assistive technologies for the deaf but also the deepening of linguistic studies. FACS is also broadly adopted in other research fields, such as Psychology and Computer Science, bringing opportunities for interdisciplinary interaction and new pathways for research in sign language, including for comparisons with oral languages, for example, referring to the lexical oral component (associated with the analysis of the word image). Finally, FACS annotation is supported by training courses, a community of annotators, and numerous examples provided by existing annotated corpora (IMOTION, 2019; EIA, 2019; EKMAN; FRIESEN, 1978).

As one of our contributions, in the present work, we propose an annotation model for facial expressions in Libras based on FACS. Our approach consisted of con-

[6] http://www.bu.edu/asllrp/SignStream/3/
[7] https://www.transana.com/
[8] https://tla.mpi.nl/tools/tla-tools/elan/

ducting a careful analysis of the Libras' facial expressions listed in Table 3.2, to encode them using FACS. However, our analysis found that there are more Libras facial expressions than those encoded by FACS. For example, there are no codes for the three possible different positions that the tongue may assume in Libras or for the action of running the tongue against the lower part of the cheek. For those cases, we proposed new codes.

According to the FACS manual, letters can indicate the intensity and position of the muscle being operated (EKMAN; FRIESEN, 1978). The letters R and L, for example, indicate whether the muscles on the right or the left are active, respectively. So, considering the existing code AU19 for tongue protrusion, we created the codes AU19R and AU19L to comprehend the three possible positions of the tongue in Libras. A similar approach was adopted in the following cases:

- the crooked mouth laterally, with AU12R indicating upward to the right, AU12L upward to the left, AU15R downward to the right, and AU15L downward to the left;
- the cheeks, with AU34R and AU34L for right and left puff cheeks, respectively;
- the closed right eyelid as AU46R and the closed left eyelid as AU46L.

In some cases, we created new codes such as AU86, AU87, and AU88, not originally present in FACS, to encode other tongue positions, such as touching the lip or the inner part of the mouth. In total was created twelve codes for the AU FACS and Libras association.

This new coding alone is not enough to cover all the facial actions that we found in Libras. However, we found that the remaining expressions could be coded as a combination of facial muscle actions, as additive AUs.

The results of this new coding, which we called FACS-Libras, are described in Table 4.4. In Table 4.4, the first column shows ours standardized nomenclature. The second column represents the mapping between the nomenclature found in the literature and the facial muscular action described in FACS (EKMAN; FRIESEN, 1978). Finally, in the third column of Table 4.4, the codes for the action units (AU) are presented. It is interesting to note that these facial expressions are not frequently observed in typical human interactions, and they are particular to sign language communication. The gray cells in Table 4.4 are the new codes created for this research.

Due to the novelty of such association, we also provide a table in the Appendix B composed with visual pictures of all the Libras' facial expressions mentioned above, including all the 53 compound facial expressions encountered in Libras.

Table 4.4 – Association between Facial Action Coding System and non-manual markers of Brazilian Sign Language

	Facial expression in Libras	**Muscular Action Description**	**Annotation Code**
Upper Face	Joined eyebrows	Inner brow raiser OR frontalis (pars medialis)	AU1
	Raised eyebrows	Inner brow raiser AND outer brow raiser OR frontalis (pars lateralis)	AU1+AU2
	Frown	Brow lowerer OR depressor glabellae, depressor supercilii, corrugator supercilii	AU4
	Left / Right eyebrow raised	Outer brow raiser Left / Right OR frontalis (pars lateralis)	AU2L / AU2R
	Wide-open eyes	Upper lid raiser OR levator palpebrae superioris, superior tarsal muscle	AU5
	Nose wrinkle	Nose wrinkler OR levator labii superioris alaeque nasi	AU9
	Slightly closed eyes	Orbicularis oculi muscle	AU42
	Closed eyes	Relaxation of Levator palpebrae superioris	AU43
	Left / Right eye closed	orbicularis oculi muscle	AU46
	Look at the speaker	Eyes positioned to look at the other person	AU69
	Direct the eyes	Eyes turn left	AU61
		Eyes turn right	AU62
		Eyes up	AU63
		Eyes down	AU64
Lower Face	Crooked mouth up	Lip corner puller OR zygomaticus major	AU12
	Crooked mouth up laterally	-	AU12R / AU12L
	Crooked mouth down	Lip corner depressor OR depressor anguli oris (also known as triangularis) AND Chin raiser OR mentalis	AU15+AU17
	Crooked mouth down laterally	-	AU15R / AU15L
	Projected lips	Lip pucker OR incisivii labii superioris and incisivii labii inferioris	AU18
	Tongue in lips position	Tongue show	AU19
	Swinging alveolar tongue		AU86
	Sibilant tongue		AU87
	Tip of the tongue touching the lips		AU88
	Tongue bite		AU16+AU19+AU22
	Contraction of the upper lip	-	AU17+AU28
	Open mouth	Lips part OR depressor labii inferioris, or relaxation of mentalis or orbicularis oris	AU25
	Contracted lips	Lip suck OR orbicularis oris	AU28
	Lower lip bite	Lip bite	AU32
	semi-open mouth (blowing)	Cheek puff	AU33
	Inflated cheeks	Cheek blow	AU34
	Only Right / Left cheek inflated	-	AU34R / AU34L
	Contracted cheeks	Cheek suck	AU35
	Clenched teeth	Lip funneler OR orbicularis oris AND Jaw clencher	AU16+AU22+AU25+AU31
	Mouth movement	Speech	AU50
	Chewing	Jaw movement	AU81
	Snap of the lips	-	AU26+AU28
	Run the tongue against the lower part of the cheek	-	AU89
Head	Neutral	-	AU0
	Head to the side	Head turn left	AU51
		Head turn right	AU52
	Head Up and down	Head up	AU53
		Head down	AU54
	Tilt to the side	Head tilt left	AU55
		Head tilt right	AU56
	Quick nod	Forward lean OR Head forward	AU57
	Brief and upward movement of the head	Tilt back OR Head back	AU58
	Balancing sideways (no)	Head shake back and forth	AU84
	Balance back and forth (yes)	Head nod up and down	AU85
	Facial occlusion	Face not visible	AU73
	Unscorable	-	AU74

The grey cells indicate the codes created from FACS to accommodate Libra's facial articulator's behaviors.

In Section 2.3.1, Table 2.2 presented the combination of action units typically associated with the expression of basic emotions and widely adopted in computational studies. In Table 4.5, we follow a similar approach to code the lexical and grammatical facial expressions in Libras.

By comparison, Libras uses 27 FACS AUs, while emotion studies use 16 FACS AUs. Only four FACS AUs of basic emotions are not involved in producing Libras' facial expressions (AUs 7, 9, 14, 20). That means, 14 FACS AUs are not usually present in studies of facial expression or even involved in existing software for action unit analysis (AU 18, 19, 22, 24, 33, 34, 34R, 35, 44, 46, 51, 52, 53, 54). In other words, by studying facial expression in sign language, we were able to transcend and extend the work on action units to those that usually are neglected.

Table 4.5 – Set of action units related with Libras' execution in discourse

Libras' facial expression function description	**Involved Action Units**
GES Question-WH	AU 4, 18, 23, 53
GES Question-YN	AU 1, 2, 4, 18, 23, 51, 52, 53
GES Doubt	AU 4, 24, 28, 44
GES Topic	AU 1, 2, 5, 18, 22, 23, 25, 26, 53, 54
GES Negation	AU 4, 15, 17, 51, 52, 54
GES Assertion	AU 53, 54
GES Condicional Clause	AU 1, 2, 53
GES Focus	AU 1, 2, 5, 18, 22, 23, 25, 26, 53, 54
GES Relative Clause	AU 1, 2
Grammatical expression of intensity	AU 4, 10, 12, 25, 26, 28, 33, 34, 35, 44, 53, 54
Grammatical expression of Homonymy	AU 10, 12, 18, 23, 25, 26, 34, 34R, 46
Grammatical expression of norm	AU 1, 2, 4, 12, 15, 16, 17, 18, 19, 28, 33, 34, 35, 51, 52, 53, 54
Total number of AU involved	27

4.5 Automatic Transcription System Overview

Considering the initial design requirements and the proposed FACS-Libras annotation model, the present work implements the automatic facial expression transcription system shown in Figure 4.3.

The system receives as input a video recording of a Libras interpreter, which will be analyzed by our classification model. The output of the FER model is the transcription of analyzed facial expressions into FACS-Libras codes. The resulting transcription is saved into a text file compatible with the ELAN tool, presented in Section 4.3.

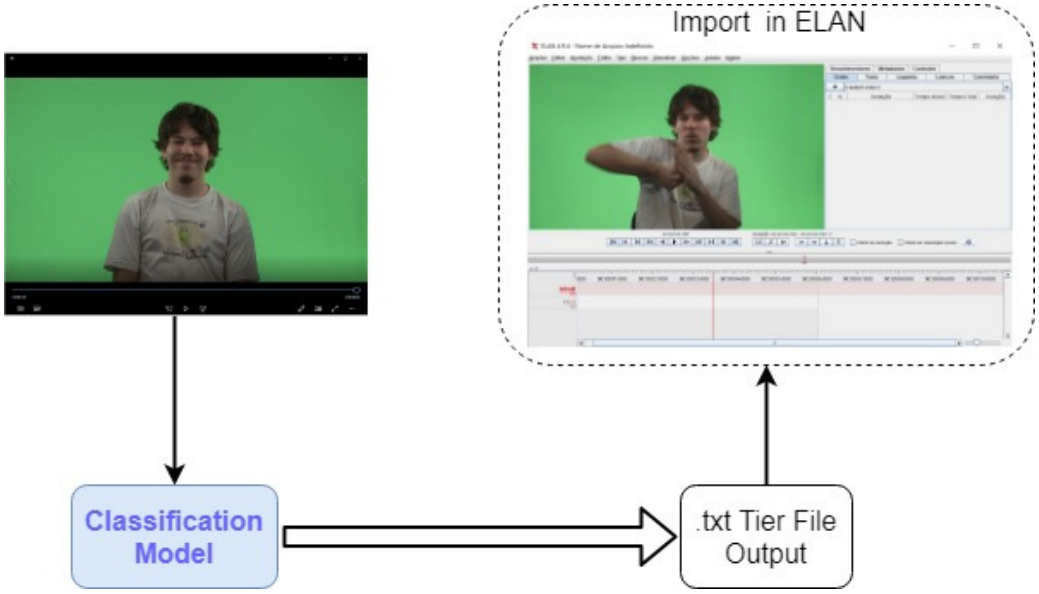


Figure 4.3 – Diagram of our complete methodology.

4.6 Concluding Remarks

This chapter presented a brief review of transcription models in the scope of Libras' facial expressions. We identified the lack of a unified annotation model, which is a crucial problem for developing an automatic classification model.

As a contribution of the present work, in this chapter, we proposed mapping between facial expressions of Libras and the Facial Action Coding System (FACS), including the creation of new codes for facial actions that were not previously mentioned by FACS.

In the next chapter, we describe the process of building two novel databases annotated with the proposed FACS-Libras annotation model.

5 Building Datasets of Libras' Facial Expressions

Well-annotated multimedia content is fundamental to the development of any recognition model based on supervised machine learning algorithms, that learn how to recognize patterns from given labeled examples.

A possible pathway to the training of FER model that recognize Libras' facial expressions would be the adoption of existing annotated corpora of Libras. In Section 5.1, we present some of them and we discuss their limitations to the present work.

Another approach would be to train the classification model with annotated datasets of facial expressions in general. However, most existing databases are encoded with emotion labels. A few of them are also labeled with FACS but, as shown in the previous chapter, the set of Libras' Facial Expressions is broader than the small set of AUs associated to the expression of archetypal emotions.

In this context, this chapter describes the building of two datasets of Libras' facial expressions, annotated with FACS-Libras, as proposed in the previous chapter. The first is HM-Libras, described in Section 5.2 and the second is SILFA, presented in Section 5.3.

The two datasets differ in at least two relevant dimensions. HM-Libras is a *non-posed* facial expression dataset, constructed from downloaded videos from the web, in the so called *"in-the-wild"* conditions. It means that the dataset captures non-acted, spontaneous behavior of the sign language interpreters, in different background and light scenarios. On the other hand, SILFA was captured in a video studio, under controlled conditions and a uniform background, asking the subjects to sign pre-defined phrases. With SILFA we guarantee the presence of all facial expressions that we need to train the model. As will be discussed in Chapter 6, both datasets were used to train our FER model.

5.1 Existing Datasets of Facial Expressions of Libras

In the literature, we could highlight only two datasets of Libras that encompass facial expressions: Grammatical facial expressions data set (FREITAS *et al.*, 2014) and RGB-D Videos in Brazilian Sign Language (REZENDE *et al.*, 2016).

In the first set of data, the grammatical facial expressions data set[1] build by Freitas *et al.* (2014), is composed of points of the face extracted using the *Microsoft Face Tracking Development Kit for Kinect for Windows*. These were obtained from ninety videos, filmed with two subjects (both hearing individuals, with dominance in Libras), five sentences of each type of grammatical facial expression of sentence, summing up to forty-five phrases. The label on each frame is the presence or absence of facial expression, without detailing the face articulators involved.

The second set of data RGB-D Videos in Brazilian Sign Language, built by Rezende *et al.* (2016), is composed of videos of ten signs: CALM, ACCUSE, ANNIHILATE, LOVE, GAIN WEIGHT, HAPPINESS, SLENDER, LUCKY, SURPRISED, and ANGRY. These ten signals, each recorded ten times, and signed by only one hearing Libras' professional interpreter, resulted in a balanced dataset with a total of 100 samples, labeled with the transcription of the signals presented in the videos.

The two datasets are available to the public for research. Nevertheless, both datasets were formed with specific characteristics, and for a particular study, so their generalization to other applications can be overly costly. Even more, when used to train a recognition system, since both datasets count with the participation of only a few subjects, it can make it bias towards a subject or even be prone to overfit to subject-identifying features. Also, neither datasets were labeled with FACS, and if they were combined, they do not comprehend most Libras' facial expressions, listed in Chapter 3. Besides, they were built with distinctive transcription words, symbols, and points of interest, which makes it challenging to reuse these sets.

Faced with the difficulties in finding a FACS labeled dataset composed by a significant number of samples of non-manual signs of Libras, we constructed a collection of videos with sentences from Libras discourse and FACS annotated. Our approach aimed at a more prominent registration of samples in Libras, whose transcription allows its analysis under various dimensions of interest for many in-depth studies of this language.

5.2 HM-Libras

Our first dataset, the Head Movement in Libras (HM-Libras) dataset was built using parts of videos of deaf individuals and sign language interpreters, downloaded from the Internet (SILVA; COSTA, 2017; SILVA *et al.*, 2020a). We downloaded videos distributed under the Creative Commons license and that are User-Generated Content (UGC).

[1] Available at https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions#

The videos were selected by using different combinations of search keywords, such as "Libras", "questions", "grammar", and "answer", in the period from 2017 to 2019. In particular, we targeted phrases with grammatical facial expressions for sentence. The HM-Libras database is composed of 80 FACS labeled videos being: 20 videos with statements, 20 videos with WH-questions, 20 videos with Yes/No-questions and 20 videos with negative sentences.

We collected videos where the person starts facing the camera to facilitate the initial face detection. These videos are not always professionally produced and often present artifacts, varying in illumination and background. The set of videos has the presence of three women and seven men. In addition to the videos, HM-Libras metadata includes a dataset matrix for each frame, composed of facial points detected using Dlib (KING, 2009). The metadata also includes the Portuguese transcription.

In summary, the dataset was built with special attention to contain samples of head movement, even more, has non-posed expressions, unscripted sentences, and in-the-wild setting. Each frame was annotated with AUs, by a single FACS coder

5.3 Sign Language Facial Action Dataset

During the production of HM-Libras, we noticed that many of the Libras' facial expressions listed in Table 3.1 were not present in the database. Aiming to train our model with the full set of facial expressions, we idealized and produced a significant video corpus composed of forty-three phrases performed by nineteen subjects. However, due to time constraints, not all those videos where annotated. Thus, we edit, annotate, and organize a subset of videos from this corpus to create the Sign Language Facial Action (SILFA) dataset.

SILFA dataset is built from the video recording of individuals signalizing twenty-three sentences in Libras. The sentences are presented in the second column of Table 5.1.

The sentences were designed to obtain targeted facial expressions of Table 3.1, including facial expressions that are not well documented in the literature (like cheeks blowing and apparent teeth), but that were observed in previous projects (TAS, 2012; SILVA; COSTA, 2017). To obtain samples on the morphological level of the facial expressions that have the function of imposing a degree of adjectivation, we define phrases that indicate changes in intensity. Thus, we choose words that already have facial expressions (e.g., beautiful), and build sentences where the degree increases and decreases (e.g., superlative very pretty, diminutive cute). In addition, some of the sentences were

Table 5.1 – Set of sentences that compose the Sign Language Facial Action Database and target facial expression classes

Task	**English Sentence**	**Portuguese Sentence**	**Sign Class**	**Sentence Class**
1	Wow! Your mother is young! How cute!	Nossa! Sua mãe é jovem! Que bonitona!	AFE	
2	Wow! Your mother is already old! So pretty!	Nossa! Sua mãe já é velhinha! Que bonitinha!	AFE	
3	Wow! Is she your mother? It is beautiful!	Nossa! Ela é sua mãe? É bonita!	AFE	
4	My girlfriend is very skinny!	Minha namorada é muito magra!	GEI	
5	My girlfriend is very fat!	Minha namorada é muito gorda!	GEI	
6	My dog is tiny!	Meu cachorro é pequenininho!	GEI	
7	My dog is huge!	Meu cachorro é grandão!	GEI	
8	My family lives very far away!	Minha família mora muito longe!	GEI	
9	My family lives very close	Minha família mora pertinho	GEI	
10	My student is very anxious about the test	Meu aluno está muito ansioso por causa da prova	GEI	
11	What is that?	O que é aquilo?	GEH	GES-WH Question
12	Who is he?	Quem é ele?	GEH	
13	Where is the toothpick?	Onde está o palito de dente?	GEN	
14	Where is my toothbrush?	Onde está minha escova de dente?	GEN	
15	Which is your magnifying glass?	Qual é a sua lupa?	GEN	
16	Have you filled your ballon yet?	Você já encheu a bexiga?		GES-Y/N Question
17	Will the two get married?	Será que os dois vão se casar?		GES-Doubt
18	Children? I do not have!	Filhos? Eu não tenho!		GES-Topic
19	He does not know a thing. It looks like he has an empty head	Ele não sabe nadinha. Parece que tem a cabeça vazia		GES-Negation
20	Now I need to work, in the future I will buy a house!	Agora eu preciso trabalhar, no futuro eu comprarei uma casa!		GES-Relative Clause
21	Why are you sad?	Por que você está triste?	AFE	GES-WH Question
22	Why are you happy?	Por que você está feliz?	AFE	
23	Why are you angry?	Por que você está bravo?	AFE	

designed to express multiple facial expressions. For instance, analyzing the task number thirteen of Table 5.1 (the sentence "Where is the toothpick?"), the sign "toothpick" has a facial expression parameter from the GEN class and the interrogative sentence has facial expression parameters from the GES class. The third column in Table 5.1 presents the class of facial expressions associated with a syntactic function. The words in which the sentences were created around were chosen from Capovilla *et al.* (2017) and BRASIL (2006) dictionaries and were further vetted by two linguistics experts.

SILFA dataset is self-contained, and it is possible to identify subpopulations by age (before 20 and over 20 years old), gender (male and female), and illumination (bright and darker). Because they are personal images, the identity of the participants is considered confidential. The sentences and images portrayed in the data are not regarded as offensive, threatening, or sensitive in any way. Due to the scope and the possibility of linguistic studies carried out in a corpus, not only of in-depth analyzes but also a comparison with other languages, the SILFA dataset creates numerous research possibilities. However, we recommend that this set should not be applied in any case that may generate unfair treatment of the deaf community, stereotyping, prejudice with malicious intent, or other undesirable harms.

SILFA dataset was created as part of constructing a large linguistic corpus of sign languages composed by a team of researchers at Federal University of ABC (UFABC)

and University of Campinas (Unicamp). The project will be updated, and new annotations instances will also be made available.

5.3.1 Participants

SILFA dataset contains video sessions of nineteen Libras users (sixteen deaf and three sign language interpreters); so far, only ten are FACS annotated (eight deaf, and two sign language interpreters). They aged between eighteen and forty-four years old, varying in physical characteristics, gender, and race, with different levels of education, all coming from the metropolitan region of São Paulo city, Brazil.

All participants gave informed consent. This research has the approval of the Research Ethics Committee (CEP) of the Federal University of ABC[2] under process number 06143719.8.0000.5594.

5.3.2 Recording Procedure

The participants were invited to enunciate twenty-three phrases in Libras, which were provided in written Portuguese. Their facial behavior was captured with high-resolution image (1440x1080 pixels) at 30 FPS (Frames per Second), by a Panasonic, AG HMC 70P, video camera. A teleprompter monitor displaying the utterance were used to facilitate its production by the bilingual deaf participants, who understand writing Portuguese[3].

The phrases were executed in sequence, two times, in one recording session. While viewing the sentence, participants sat in a chair positioned in front of the camera. They were provided with a Libras interpreter in the room. Figure 5.1 brings the adopted configuration for the recording.

No type of study was carried out with the participants after the data capture was performed.

5.3.3 Manual FACS Annotation

The video materials were transcribed with the aid of the ELAN software (BEREZ, 2007). ELAN software allowed the annotation of the facial expressions present in each sentence, as well as the signs and the Portuguese translation through the use of

[2] To be clear, the Federal University of ABC is the institution of the co-supervisor of this research.

[3] Since the oral language and the sign language constitute different languages and channels for the transmission and reception of linguistic information, individuals who have Libras as their first language may not have learned the written Portuguese language form (SKLIAR, 1998). We highlight that our subjects are capable of reading written Portuguese.

Figure 5.1 – Studio setting used for recording SILFA.

tiers. At this stage, the total duration of the utterance was considered as the interval between the output from a neutral facial expression until the return to this position.

AU presence was coded for each video frame. FACS coding was performed by a single FACS coder and Libras' facial expressions were transcribed by one deaf and one hearing individual, both Libras speakers. It took roughly five months of our project for this annotation process. Metadata consists of manually annotated AUs, Portuguese transcription, Libras taxonomy of facial expressions, and facial landmarks.

5.3.4 SILFA Qualitative Analysis

In this section, we present a qualitative analysis of the resulting annotated SILFA database. In $15,6\%$ of the sentences in which the intensity degree increased (superlative), it was observed the presence of the same face, a combination of AUs 1+2+34 (raised eyebrows and inflated cheeks). Participants consistently used AUs 4+35 (frown and contracted cheeks) on $9,4\%$ of phrases where the intensity decreased (diminutive), which indicates a pattern of those AUs to represent determined intensity facial expression. For that reason, it was adopted in this research the nomenclature "intensity face", since they were used to express adjective intensity and are defined by a subset of AUs[4]. As can be seen in Figure 5.2, in images (A) and (B) we illustrate the facial expression "raised eyebrows and inflated cheeks", which is a combination of AUs 1+2+34. Also, in

[4] This terminology will be applied later on.

Figure 5.2 – The "intensity face". Images (A) and (B) are examples of facial expressions associated with the superlative kind of intensity. Images (C) and (D) are examples of facial expressions related to the diminutive type of intensity.

images (C) and (D), we illustrate the facial expression "Frown and contracted cheeks", a combination of AUs 4+35. In summary, the results of this experiment is a first step towards demonstrate the consistency of AU production of both intensity face across participants of different backgrounds and age groups. Additionally, we had asked, previously at recording, that the subjects employed facial expression in their interpretation.

The analysis indicates that in sentences from the GEI class, almost every participant executed the sign by pattern[5]. In a few cases, they change the intended manual sign, but they do the standard facial expression. However, in sentences of GEH class, the subjects oscillated between changing the manual sign and not presenting facial expression. For instance, in the sentence "I am sorry. Tomorrow I can not", the intended sign is "can not" (in Portuguese gloss, "NÃO-PODER"). According to Capovilla *et al.* (2017), it was stipulated that the standard is the NMM "head shake" combined with a manual sign of "occupied" (in Portuguese gloss, "OCUPAD@"). But, two subjects made the same manual sign as "power" combined with "head shake", as illustrated in Figure 5.3. One person did a completely different manual sign, in which the sign "to leave" was used combined with "head shake", and five participants made the pattern sign. In phrases "I am looking for a motel" and "My teacher is a lawyer", similar behavior was observed.

Figure 5.3 – Libras' signs for "Not" and "Can" in the performance of the sentence "I am sorry. Tomorrow I can not". Images from (SILVA *et al.*, 2020b).

We discuss the expressions in the first phrase, "I am looking for a motel". From the ten participants analyzed, only four had one of their cheeks inflated combined with the hand sign in their performance. Such a combination form our standard definition for

[5] In this study, the Libras' sign by pattern or the standard version of a Libras sign it is considered as described in Capovilla *et al.* (2017).

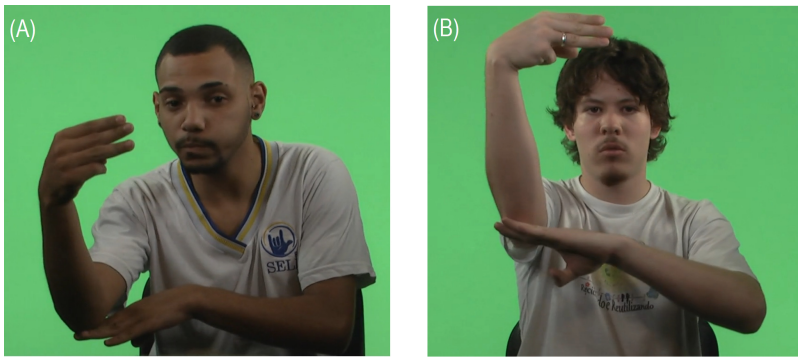


Figure 5.4 – Illustration of the Libras' sign "Motel". In (A) is portrayed the standard form, and in (B) is the modified form found in (SILVA *et al.*, 2020b).

the sign. The Libras signs "Motel" and " Hotel", in their standard form, are differentiated only by the NMM of the cheek stretched with the tongue. We observed that six of the participants changed the hand configuration of the sign as a strategy to distinguish the signs, performing "Motel" with the hand configuration in the letter "M" (motel initial) of the Libras fingerspelled words, as illustrated in Figure 5.4. Thus, the distinctive feature between the signs passed from facial expression to the hand.

Regarding the behavior of the face in the sentence "My teacher is a lawyer", three subjects performed the Libras' sign "Lawyer" by the defined pattern, six performed the manual expression without NMM, and two performed the sign for "justice" as illustrated in Figure 5.5.

Note that in our data, we identified cases in which different signs produced different manual signs and absence or modification of NMM to refer to the same event. In the Libras' sign "can not", we observed a recurrent configuration as projected lips accompanying the negative sentence. Interestingly, the performance of these mouth action is synchronized with the headshake movement. The data discussed here can illustrate a trend of a strong presence of GEI use and variation in the employment of GEH, where in some cases, they have been dropped.

Figure 5.5 – Illustration of the Libras' sign "Lawyer". Image (A) presents the expected form for the sign, and in (B) is the sign "Justice". Images from (SILVA *et al.*, 2020b).

5.4 Concluding Remarks

In this chapter, we detailed the building of two datasets used to train our FER model. By describing and exposing both data sets, their differences and characteristics can indicate their possibilities of applications. Some results from the manual annotation of the SILFA database show some novel observations from the capture of such rich data demonstrating its potential to contribute to linguistic studies in Libras.

The next chapter presents our method that uses both HM-Libras and SILFA in training and testing.

6 Libras' Facial Expression Recognition

In this chapter, we introduce the modules that compose our Libras' facial expression recognition system. In Figure 6.1, we revisit the general framework of FER systems, as previously presented in Chapter 2. The figure highlights the main modules of our system, and also provides an overview of the chapter organization.

In Sections 6.1 and 6.2, we describe the methodology we adopted to model our pre-processing stage. Similarly, in Section 6.3 , we describe the neural network architectures we considered in the present work. As illustrated by Figure 6.1, the integration of these modules forms our FER system that processes input videos of a sign language interpreter and outputs an ELAN annotation file.

The codes for the frameworks described in this chapter are available at: <https://github.com/SrtaEmely/FacialActionLibras>

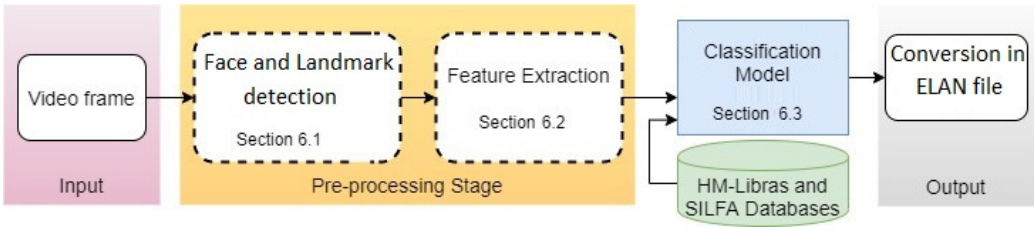


Figure 6.1 – Steps applied in the proposed Libras' FER method. Image created for the study itself.

6.1 Face and Landmark Detection

In the pre-processing stage we adopt a machine learning technique to find the face in the image. Detecting the position of a face on an image may be a challenging task due to the rigid (scale, rotation, and translation) and non-rigid (such as facial expression variation) face deformation. A robust algorithm capable of learning not only the position of the face but also indicating the position of facial elements is ideal. As we mentioned before, a face detector with great accuracy is a regression-based method constructed by a classic HOG feature combined with a linear classifier, an image pyramid, and sliding window detection scheme (Haar-Cascade features) (VIOLA; JONES, 2004; KAZEMI; SULLIVAN, 2014). We applied OpenCV's (BRADSKI; KAEHLER, 2000) built-in Haar cascades features.

After the detection of the face in the image, a fixed bounding box is drawn

around and cropped from the picture. Then, the new image containing only the face is resized into 96 by 96 pixels. Although many databases are composed of higher-resolution images, tests suggest that decreasing the resolution does not significantly impact accuracy, and has the advantage of increasing the computation speed of the network (MOLLAHOSSEINI *et al.*, 2016). Based on the advantages of a hybrid face detection approach, we also extract the landmarks. Using the pre-trained facial landmark detector inside the Dlib library, we extracted 68 landmarks localized on the face placed alongside the ears, chin, nose, eyes, eyebrows and, mouth (KAZEMI; SULLIVAN, 2014). This method estimates facial landmark positions directly from the pixel intensities themselves and was build upon an ensemble of regression trees. Figure 6.2 shows the steps of the face detection and the positions of the landmarks on the face.

6.2 Feature Extraction

Raw data, as it is extracted from the camera, can be directly used as the input characteristic vector of deep learning models. However, in this work, we hypothesized that pre-processing steps for feature extraction, would help the network to learn improved discriminative dimensions, resulting in better AU recognition rates. As a consequence, our feature extraction methodology was established based on the results of systematic experiments described in detail in the following chapter (Section 7.1.3).

Our studies showed that a higher recognition accuracy was obtained when the input facial image was cropped in two regions: (1) the lower portion of the face, comprehending the chin, the mouth, and the nose; and (2) the upper portion of the face comprehending the forehead, the eyebrows, and the eyes. Note in Figure 6.3, there is an overlapping between ROIs. Further, each of the sections has dimensions 96 by 60 pixels. We define the upper segmentation by start counting from the top down. At the same time, the lower segmentation starts counting from the bottom up.

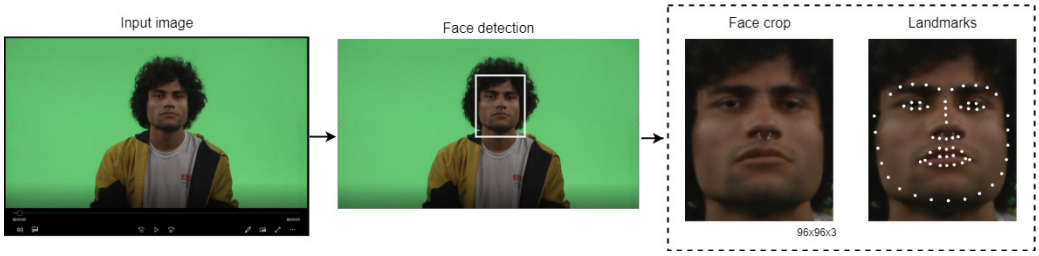


Figure 6.2 – Diagram of face detection step. Given an example input image, the face detection is obtained by OpenCV (BRADSKI; KAEHLER, 2000) and Dlib (KAZEMI; SULLIVAN, 2014) implementation. We extracted the copped face and the positions of the landmarks. Image created for the study itself.

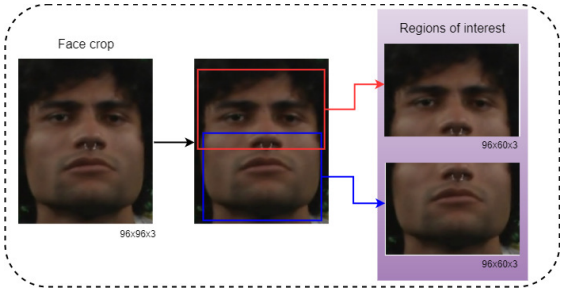


Figure 6.3 – Diagram of the region of interest stage. From the face cropped image, we segmented into upper and lower regions of interest. To embrace the whole movements of the facial muscle, we let an overlapping between the regions occur. Image created for the study itself.

In addition, the experiments showed that the system performed better when we embedded geometric facial features into the input image. To obtain the geometric features we, first, calculate the distance between the middle points in the lid tighteners in the eyes to indicate if the eyes are open or closed. Likewise, for the mouth, we calculate the distance between the midpoints of the upper and the lower lips. Each of these measurements was then converted into a single gray pixel. In other words, we compose vectors with the face points p_i, $i = 1, \ldots, 68$ and the distance measures $d_2(p_j, p_k)$ with $(j, k) \in \{(3, 13), (17, 21), (21, 22), (22, 26), (38, 40), (43, 47), (48, 54), (51, 57), (62, 66)\}$. Following, these values are normalized to the range $0 - 1$ which creates a vector of abstract color components, and then are encoded as gray levels (that vary from 0 to 1, allowing all range of real numbers inside this set). By expressing the intensities of pixels from the obtained normalized values as gray levels, it is defined the range from 0 or black as in total absence, 1 or white as in full presence, and any fractional values in between as of shades of gray. The details can be viewed in Figure 6.4. Finally, the gray vector is concatenated and added as columns in their corresponding facial region images. Figure 6.5 shows our complete pre-processing stage.

6.3 Classification Model

Our approach to the design of a Libras' facial expression classification model consisted of conducting experiments with different neural network architectures, and evaluating their performance according to predefined metrics.

In particular, in this section we highlight three different architectures.

As discussed in Chapter 2, the core of many FER models is a convolutional neural network. Based on that, we implemented a CNN as our baseline model. Following, we evaluated an architecture that combines the ability of CNNs to extract relevant features from images, with the ability of Long Short-Term Memory (LSTM) networks to

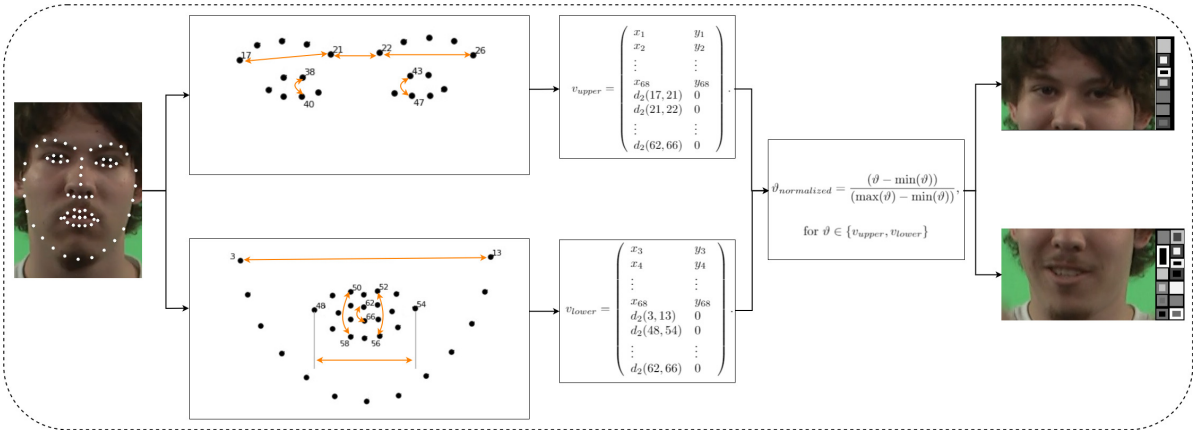


Figure 6.4 – Diagram of geometric features extraction. From the positions of the landmarks, we calculated the distances highlighted in lines colored orange. With the obtained values, we created two vectors v_{upper}, v_{lower} corresponding to each region of interest. The next step is a unity-based normalization. Later, the gray level vectors are concatenated to their respective region of interest. Image created for the study itself.

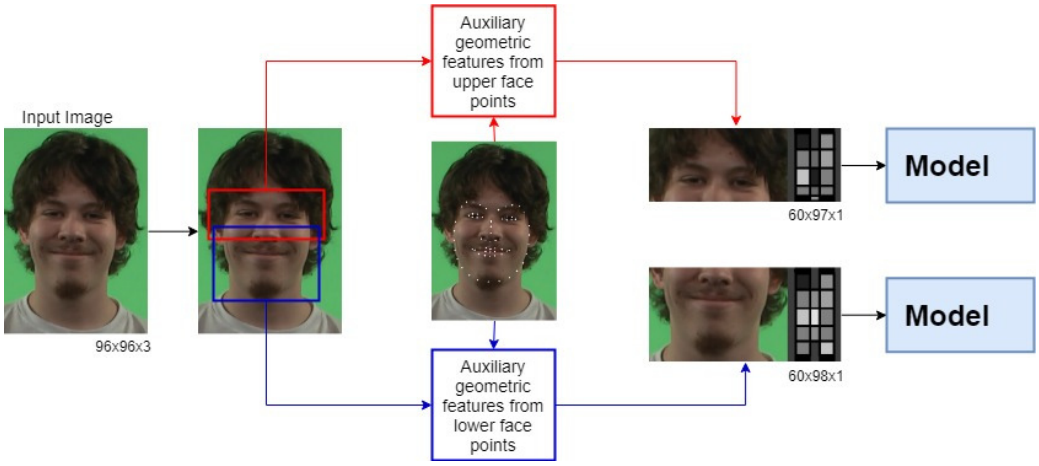


Figure 6.5 – Diagram of full pre-processing stage. Image created for the study itself.

model the correlation of frames in temporal sequences. Finally, we explored and evaluated SqueezeNet, a compressed CNN architecture with a successful trajectory in object detection and recognition.

In Chapter 7, we detail and discuss the results of our experiments. Among the three architectures, SqueezeNet obtained the best results.

In the following sections, we describe the architecture details of the three classification models.

6.3.1 CNN

As our baseline model, we implemented a shallow CNN with five hidden layers. The design is based on the information that a shallow CNN is already able to learn

high-level discriminatory features to design our network (MOLLAHOSSEINI *et al.*, 2016; PRAMERDORFER; KAMPEL, 2016). Our model consists of a CNN in which the first layer is a convolutional layer, the second layer is a max pooling layer, the third layer is a convolutional layer, the fourth layer is another max pooling layer, the fifth layer is a flatten layer, the six and the last layer are dense layers. The activation functions are all set to ReLU (Rectified Linear Functions). The configuration model and other details are shown in Figure 6.6.

In Figure 6.6, the output of the fully connected layers are regarded as the encoded features. The classifier is the combination of the fully connected layer (L6) and the softmax output layer (L8). The feature extractor for the input image is the whole structure with the exception for the classifier. From the convolutional layer (L1) to the Max pooling layer (L4) is defined as the Convolutional Feature Extractor (CFE) while the Fully Connected Feature Extractor (FCFE) is defined as from flatten layer (L5) to fully connected layer (L7).

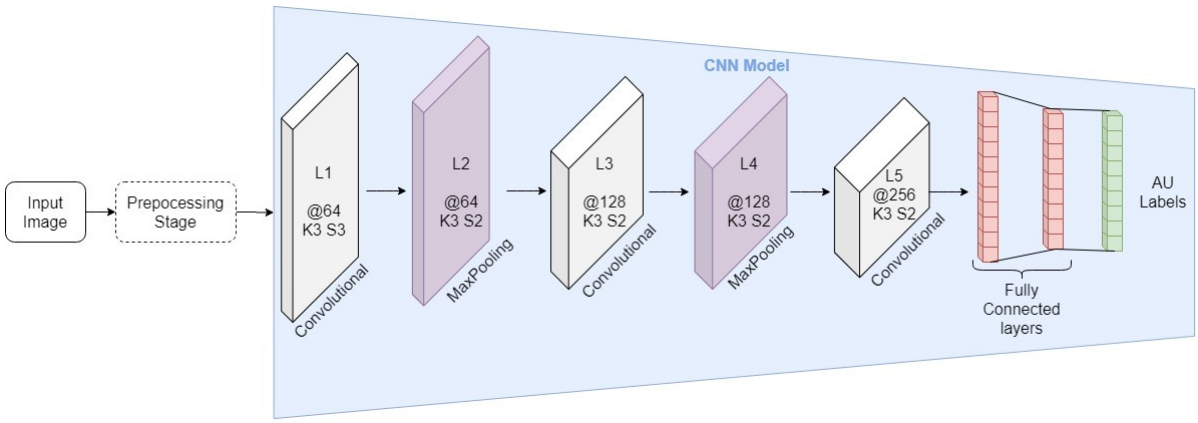


Figure 6.6 – CNN architecture. @, K and S denote number of filters, kernel size and stride, respectively. Image created for this study.

6.3.2 CNN+LSTM

Since an AU is an observable event throughout the time, many works hypothesize that the recognition of facial expressions can be improved by the knowledge of previous states. Inspired by the works of Li *et al.* (2017a), Mei *et al.* (2018) and Chu *et al.* (2019), we extended our system to address temporal context by designing a combination of both CNN and LSTM, to fuse static features with temporal cues.

More specifically, we propose a standard CNN with three convolutional layers alternated by two max-pooling layers. The convolutional layers are composed by kernels of size three and stride one. The first two convolutional layers have 32 filters and the last one

has 64 filters. The max-pooling layers have a stride[1] of size two. All activation functions are set to ReLU. The last layer is the region pooling layer. We model the correlations between spatial and temporal cues by adding a fusion layer. This fusion layer is a concatenation of feature maps, made to get regional features. So, from the CNN, we have a pooling layer with 30 filters for the upper part of the face and 50 filters for the lower part of the face. These feature maps, are fed into stacks of LSTMs to fuse temporal dependency. We combine two frames of images as a sequence into the LSTM. Essentially, the timestep is the number of units used to predict the future steps, in our case, the next frame. This number also relates to the amount of data per batch. Hence when we register the time step as two, we can consider two sets of data equals to two batch-size in which are taken to predict the next frame. Therefore, 2 groups of the batch-size form around 3 seconds of a video (an estimation for the performance of facial expression reach the apex). Then several stacks of LSTMs are used to capture facial action temporal dependency. Finally, the outputs of LSTMs are aggregated into a dense layer to perform multi-label learning. In Figure 6.7 we present a scheme of CNN+LSTM network architecture with the LSTM cell.

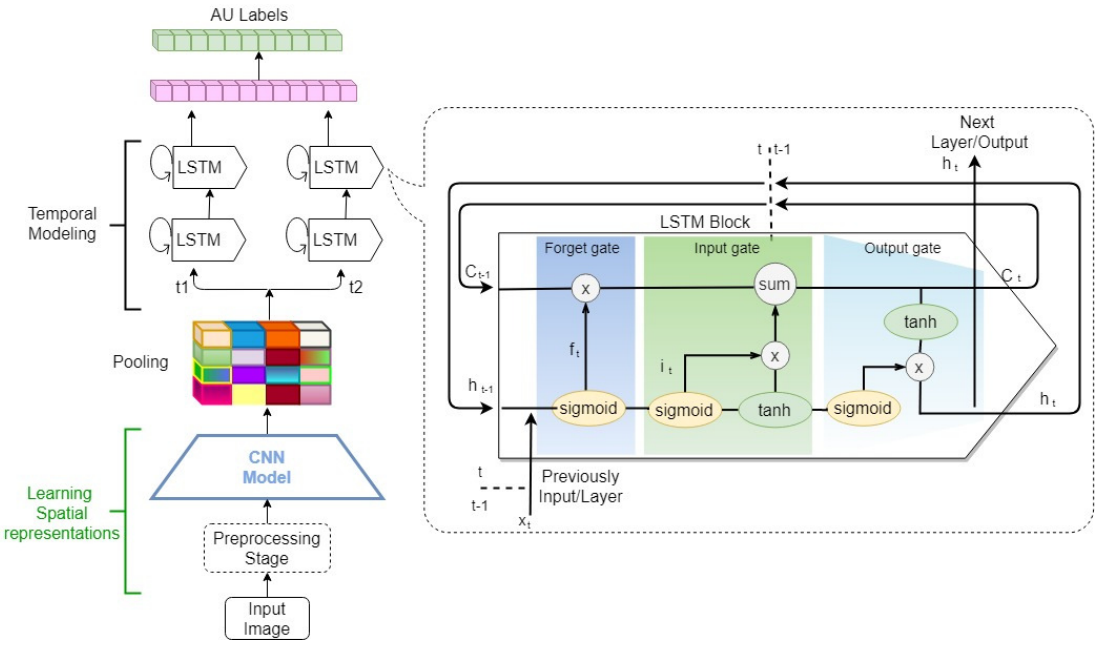


Figure 6.7 – CNN+LSTM architecture with the architecture of the LSTM cell showing the repeating module that contains four interacting layers. The illustration was inspired from (CHU *et al.*, 2019) and (FAN *et al.*, 2020).

[1] Stride is an implementation component of neural networks which set a parameter that modifies the amount movement over the position of values in the matrix to create a filter. The size of the filter affects the encoded output and is often set as an integer. For example, given a matrix and our filter with stride equals 1, then the filter will move one position at the time throughout the whole matrix (DESHPANDE, 2016).

6.3.3 SqueezeNet

Based on a CNN architecture, the SqueezeNet brings advantages by allowing information to flow through its layers resulting in a powerful ability[2] to generalize to unseen data while maintaining its high accuracy rate. The design strategy adopted by Iandola *et al.* (2016) for SqueezeNet, was to introduce the **Fire module**, a new building block composed of a convolution layer which has only 1×1 filters, and called *squeeze*, feeding into a layer that has a mix of 1×1 and 3×3 convolution filters, called an *expand layer*. The use of fire modules helps to limit the number of input channels to the 3×3 filters, making fewer parameters in tunable dimensions (hyperparameters).

The SqueezeNet architecture begins with a standalone convolution layer followed by a max-pooling layer. Next, three fire modules are stacked, followed by a max-pooling layer, in sequence, four fire modules are stacked followed by a max-pooling layer, and another two fire modules ending with a final convolutional layer. The max-pooling layer applied is composed with a stride of 2. Also, the ReLU is applied to activations from squeeze and expand layers. Since the output activations of the Fire modules need the same height and width from 1×1 and 3×3 filters, it was added a $1-$pixel border of zero-padding in the input data to 3×3 filters of expand modules (IANDOLA *et al.*, 2016). In Figure 6.8, we present the SqueezeNet architecture and the diagram for a fire module.

[2] To be clear, due to the severe dimensionality reduction from the bypass connections in SquezeNet, the representational bottleneck introduced by squeeze layers helps to alleviate the number of parameters in the network while still allowing it to be deep and represent many feature maps.

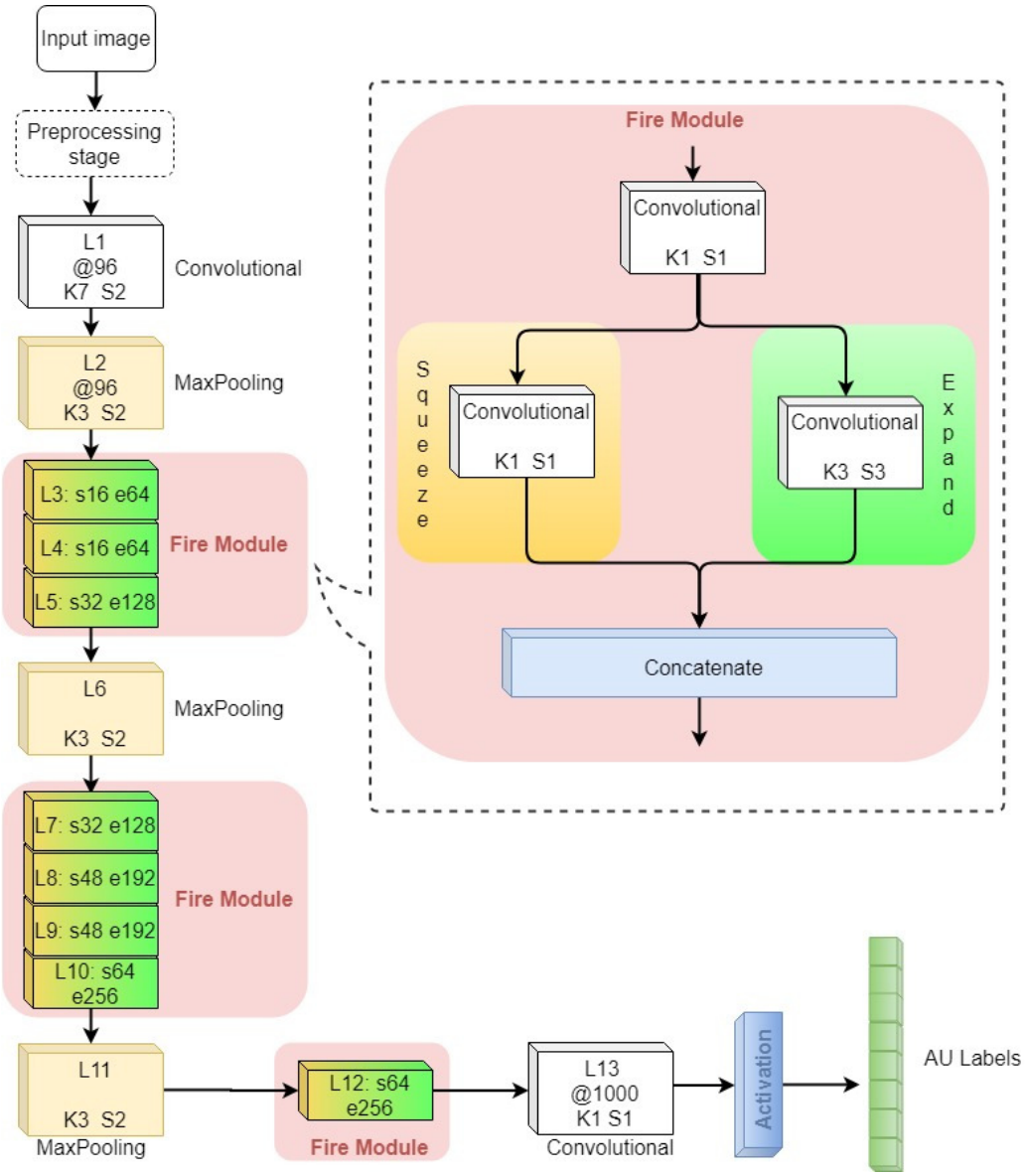


Figure 6.8 – SqueezeNet architecture. @,K,S,e,s denote the number of filters, the kernel size, the stride, the expand filters, and the squeeze filters, respectively. The illustration was inspired from (KATSIOS, 2019).

6.4 Concluding Remarks

This chapter provided an overview of our Libras' facial expression recognition system, given the reference framework we described in Chapter 2. Also, by detailing the chosen networks architectures, their advantages and particularities to our FER model we detailed our complete approach.

Each module of the final framework was designed according to the results of experiments. The experiments and their results are shown in Chapter 7.

7 Experiments and Results

In this chapter, Section 7.1 highlights three experiments that helped to define the current state of our Libras' facial expression recognition system. In Section 7.1.3, we show the results of a first experiment, that helped to define the design our preprocessing stage. Following, in Section 7.1.4, we show the results of a comparative analysis among neural network classifiers, with focus on the three architectures we presented in last chapter. The third experiment is presented in Section 7.1.5, and consists of a cross-database test.

The chapter also present a comprehensive analysis of the use of our final FER system as an end-to-end facial expression annotation system for Libras. Section 7.3 compares the annotation results of our system with the annotation provided by two trained annotators. We then discuss the open challenges and the benefits presented by our system.

7.1 Experiments

This section starts with a description of the metrics we adopted to evaluate the results our experiments, followed by the experiment's implementation details. Later, we describe three computational experiments. The first experiment helped to establish the feature extraction preprocessing method we described in Section 6.2. Second, in Section 7.1.4, we compare the CNN, CNN+LSTM and SqueezeNet architectures, with other architectures. The results of the second experiment justifies our focus on these three architectures in the following experiment, in which we performed tests across databases to evaluate our target architectures. The section ends with a comprehensive discussion of our results.

7.1.1 Metrics

The performance of AU recognition was evaluated on F1 frame-basic metric, average accuracy, precision, and recall (also called sensitivity). F1 score is the harmonic mean of precision and recall and is widely used in AU recognition (SILVA *et al.*, 2020a; KOELSTRA *et al.*, 2010; ZHAO *et al.*, 2018; CHU *et al.*, 2017). When samples are not balanced, the F1 score can describe better the performance of an algorithm (SILVA *et al.*, 2020a; LI *et al.*, 2018).

7.1.2 Implementation Details

All architectures were evaluated based on Keras implementation (CHOLLET *et al.*, 2018). For the experiments we traind the networks architectures from scratch without any pre-trained weights. We train every architecture for up to 300 epochs and a fixed mini-batch size of 128 samples. Both models were initialized with a learning rate of 0.01, optimizing the cross-entropy loss using stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.001. Simard *et al.* (2003) have shown that if the data is augmented in a reasonable way, the model can perform better. For training data augmentation we used horizontal mirroring, randomly rotations and two types of shift and zoom transformations. These are applied indiscriminately in each epoch creating twice the amount of data.

Due to the gestures in SL, we must take into account the frequent occlusion of the face. When that happens we decided to drop the frame, labeling it with AU code 73. That happened in $3,4\%$ of HM-Libras and 3% of SILFA datasets. In DISFA dataset, it was dropped 0.2% of samples mostly due to partial occlusions or extreme head rotations.

All experiments were performed on a PC with one NVIDIA GTX 1070 GPU. It took roughly 160 hours to train each network until the convergence, that happened around 250 epochs.

7.1.3 Experiment 1: Ablation Study

In this experiment, we trained the CNN described in Section 6.3.1 with three different feature extraction strategies: (1) the whole face image as extracted from the entry; (2) the face image separates into regions, the upper and lower facial parts; and (3) the face image split into regions with geometric features as described in Section 6.2.

In the first type of entry, the detected face image was resized into $96 \times 96 \times 3$. The second input type, consisted of cropping the detected face image into two regions. The upper part concerns eyebrows and eyes, with a size of $60 \times 96 \times 3$. This segmented image was made when selecting the face image by starting from the top down. Whereas, the lower part concerns cheeks, mouth, and chin, which also has a size of $60 \times 96 \times 3$. This second segmented image was made when chopping the face image by starting from the bottom up. There are overlapping between those two face region images. The last entry type is described in Section 6.2.

Figure 7.1, illustrates each type of input. To facilitate the understanding, we named each type of pre-process as: *No Preprocessing* (Figure 7.1(A)), with *Region of Interest* (Figure 7.1(B)), and with *Region of Interest and Geometrical Features* (Figure 7.1(C)). Table 7.1, shows the average accuracy and F1 measure for AU recognition based

in each type of preprocessing.

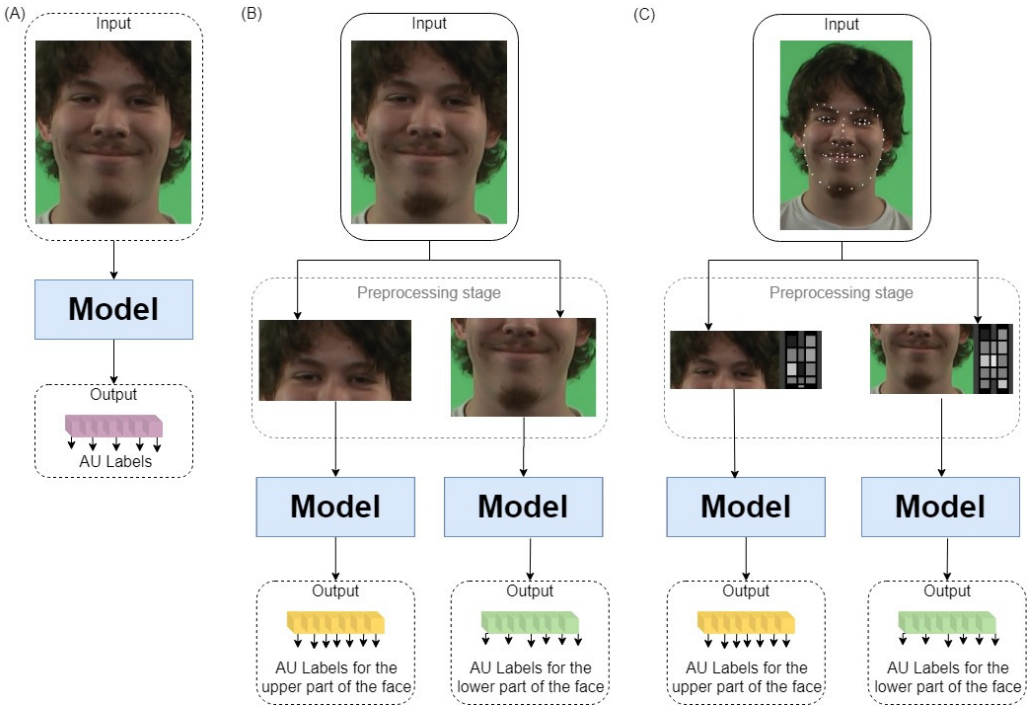


Figure 7.1 – The three pre-process techniques employed in experiment one. Image (A) shows the *No Preprocessing* system. Image (B) shown each step in the *Region of Interest* scheme. Lastly, presented in the image (C) is the *Region of Interest and Geometrical Features* design. Image created for the study itself.

Table 7.1 – Network performance evaluation according to the type of preprocessing scheme

	Avg acc	Avg F1
No Preprocessing	0.4992	0.4834
Region of Interest	0.6797	0.6913
Region of Interest and Geometrical Features	0.7628	0.7495

According to Table 7.1, when we compared between no preprocessing stage and region of interest with geometrical features, there was a significant improvement in the classification rate. These results can be attributed to the focus on the image information area, which cause to the network to benefit more. Also, larger images produce large feature maps, which has a negative impact on high-level features. In the following experiments, we adopted the feature extraction method described in Section 6.2. This choice provides accurate guidance for the network.

7.1.4 Experiment 2: Comparing Neural Network Architectures for Classification

In order to verify the universality of the proposed AU detector framework, we also conducted comparative experiments on two well-known networks in AU-based

FE recognition: AlexNet and VGG-16. This experiment aims to determine which of the architectures described in Section 6.3 perform better than AlexNet and VGG, which uses entire face images to learn.

We evaluated the proposed architectures performing experiments with the databases: Extended Cohn-Kanade dataset (CK+), DISFA (Denver Intensity of Spontaneous Facial Expressions), the HM-Libras database and the SILFA database. We choose CK+ and DISFA since they have AU annotations and are standard datasets for AU detection, and to the best of our knowledge, HM-Libras and SILFA are the only Libras database with AU annotations. To ensure that the subjects are mutually exclusive in the train/test split sets, we adopted a form of the 3-fold partition by forming a group with CK+, DISFA, HM-Libras, and SILFA for training and using only a percentage of HM-Libras and SILFA for a test set. That way, we are able to demonstrate the effectiveness of the proposed model to the Libras application since extensive experiments have been conducted on our already described networks using a subject independent and cross-database approach. The overall performance of our algorithms is described by the average F1 score displayed in Table 7.2.

Analyzing Table 7.2 we can make a few observations. First, without the region constraint on AU and geometrical landmark detection, the baselines methods perform the worst. It can be seen in Table 7.2 that our proposed framework with SqueezeNet outperforms all other architectures. Compared with other works, the proposed framework with SqueezeNet achieves competitive performance with more AU labels, which demonstrates the effectiveness of our proposed region and geometrical features learning.

Table 7.2 – Comparative performance between networks

Architecture	**preprocessing**	**Metrics**			
		Precision	**Recall**	**Acuraccy**	**F1**
AlexNet	upper face	0.7199	0.6226	0.6620	0.6677
	lower face	0.6827	0.5719	0.5402	0.6224
	Avg.	0.70	0.59	0.60	0.64
VGG-16	upper face	0.5123	0.5098	0.5322	0.5110
	lower face	0.4732	0.4698	0.4082	0.4714
	Avg.	0.54	0.53	0.47	0.49
CNN	upper face	0.8564	0.6697	0.7999	0.7516
	lower face	0.8187	0.6670	0.7257	0.7351
	Avg.	0.83	0.66	0.76	0.74
CNN+LSTM	upper face	0.7789	0.6349	0.71	0.6995
	lower face	0.7619	0.5575	0.6898	0.6438
	Avg.	0.77	0.54	0.69	0.67
SqueezeNet	upper face	*0.8456*	*0.7734*	*0.8091*	*0.8078*
	lower face	*0.8640*	*0.6592*	*0.7435*	*0.7478*
	Avg.	**0.85**	**0.71**	**0.77**	**0.78**

The comparison between CNN and CNN+LSTM shows a decrease in average accuracy. Our belief is that the reason for the overall performance drop is that the short time steps window to the LSTM was not capable of extracting the facial action. According to Ma *et al.* (2019), some AUs have a drastically different variation time, and if the temporal length of AU duration is short then the CNN+LSTM model could not observe such actions. Further experimentation was not made on CNN+LSTM.

Our task was to detect whether the AUs are active (present), which is a multi-label binary classification problem. To conduct AU-level analysis, we show the AU occurrence rates in the same training scheme averaged mentioned above in Table 7.3.

We can observe that the occurrence rates of AUs 4, 22, 26, 27, 34, 43, and 45 are much higher than those of other AUs for our HM-Libras and SILFA test set. It can be observed that the proposed framework with SqueezeNet model significantly outperforms the AlexNet baseline. By comparison, we can clearly see in Tables 7.2 and 7.3 that the average f1 measure for the AU combination made in our multi-label AU detection slightly outperforms the single AU recognition approach in most AUs. Additionally to the experimental values given in the tables above, we also provide in Figure 7.2 the confusion matrix. By observing its color disposition, it is possible to get insights about which classes our model recognizes well and which it confuses. Figure 7.2 makes it clear that the model confuses similar AUs like, for example, 4+18 and 4+33, or 15 and 15+17. Another remark is about the AU0 that receives high classification rates. Further analyses will be made later in the chapter.

Table 7.3 – Comparative performance between networks

Model	**FACS - AUs**																	
Architcture	**0**	**1**	**2**	**4**	**5**	**9**	**10**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**22**	**23**
AlexNet	0.64	-	-	-	-	-	-	-	0.12	-	-	-	-	-	-	0.01	0.46	
CNN	0.91	0.10	0.36	0.70	0.24	0.5	0.67	0.30	0.30	0.13	0.09	0.76	0.48	0.44	-	0.38	0.29	0.02
SqueezeNet	0.83	0.39	0.5	0.75	0.39	0.61	0.53	0.52	0.61	0.61	0.49	0.58	0.54	0.44		0.61	0.77	0.52
Model	**FACS - AUs**																**Avg.**	
Architcture	**24**	**25**	**26**	**27**	**28**	**33**	**34**	**35**	**43**	**45**	**46**	**61**	**62**	**63**	**64**	**73**		
AlexNet	0.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.58	
CNN	0.54	0.51	0.64	0.52	0.32	0.01	0.08	0.01	0.40	0.31	-	0.02	0.4		0.07	0.22	0.69	
SqueezeNet	0.44	0.56	0.70	0.86	0.37	0.25	0.67	0.22	0.69	0.94	0.59	0.05	0.27	0.44	0.82	0.26	0.73	

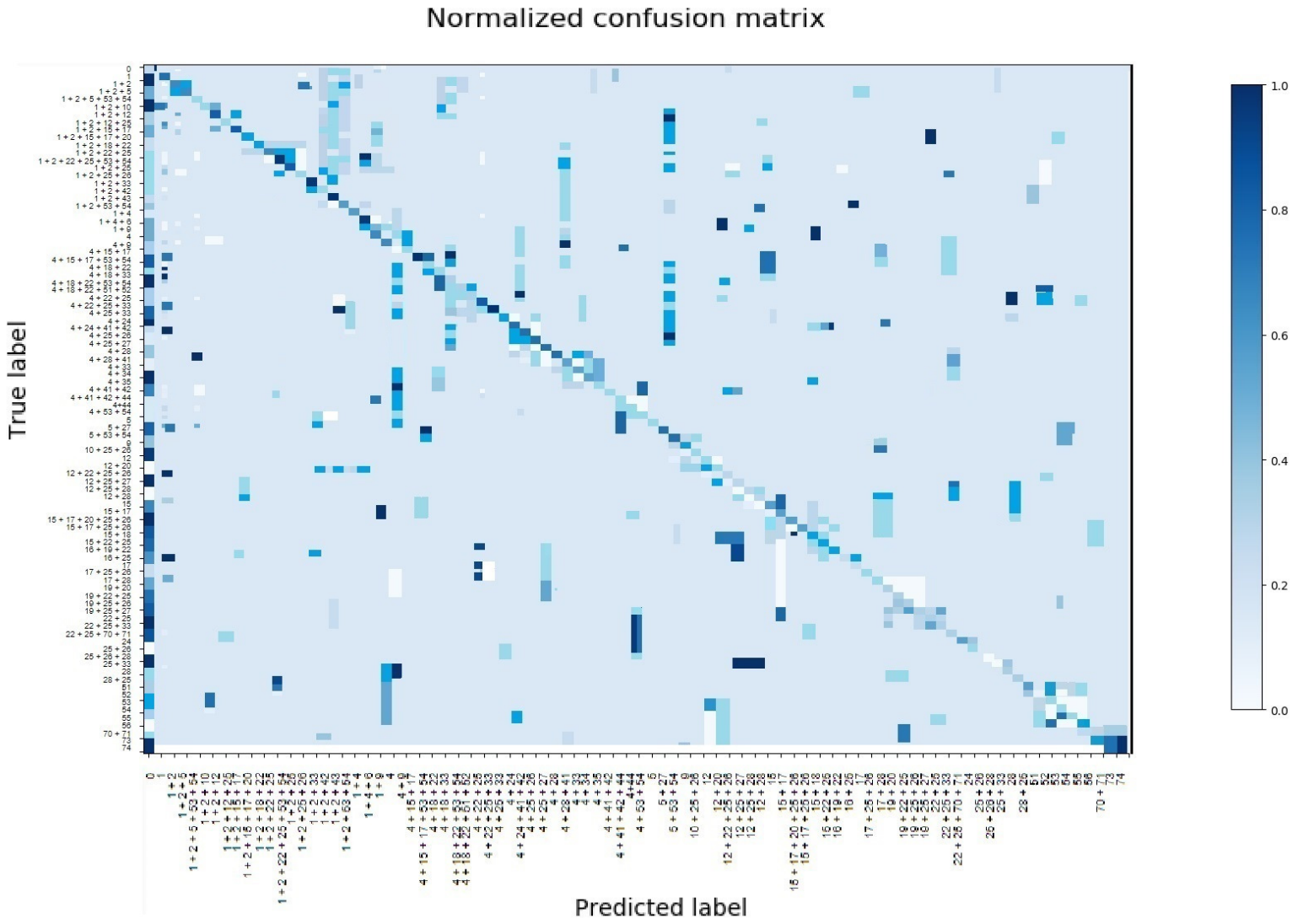


Figure 7.2 – Confusion matrix for 90 classes obtained on experiment 2 from the SqueezeNet model. In this normalized confusion matrix, the dark blue color denotes higher numbers, and the white color corresponds to the empty miss-classifications containing zeros. Note that it is main diagonal is quite distinct; in most cases, the predicted label coincides with the ground truth. Source: the research itself.

7.1.5 Experiment 3: Cross-Database Testing

After performing within-dataset tests on a combined set CK+, DISFA, HM-Libras, and SILFA, we conduct a cross-dataset test on the large-scale DISFA when training on HM-Libras and SILFA. That was done so we could compare our approach with other works in the field. These works include Han et al. (2018) (HAN *et al.*, 2018), AlexNet (KRIZHEVSKY *et al.*, 2012), Li et al. (2017) (LI *et al.*, 2017a), EAC-Net (LI *et al.*, 2018), Mei et al. (2018) (MEI *et al.*, 2018), Sankaran et al.(2020) (SANKARAN *et al.*, 2020), ARL (SHAO *et al.*, 2019), and STRAL (SHAO *et al.*, 2020b) which report results on a subset of the 12 action units. Although there is a domain gap between the test set (DISFA) and our set of training data (CK+, HM-Libras, and SILFA), the SqueezeNet model achieve good performance on DISFA. In Table 7.4, we present a comparison between our approach and the state of the art. The experiments conducted on this dataset followed the same division strategy of the previous papers, which is the subject-exclusive 3-fold cross-validation setting. Note that most of these approaches use outside training data or additional facial landmark labels, while our method differs in training by using HM-Libras and SILFA database with more AU labels. The results are shown in Table 7.4. Such experiment was done so that we could compare our approach with other works in the field and gain further understanding regarding the learned features of the different AUs. Moreover, we can see that SqueezeNet model outperformed the state-of-art in terms of both F1-frame and accuracy metrics, which demonstrates the better generalization ability of our framework.

The values presented in Table 7.4 for the proposed SqueezeNet-based framework are close to the state-of-the-art metrics for some AU labels on DISFA. We see that our novel framework can adapt to various AUs with different variations, sizes, and non-rigid transformations for the AU detection task. Note that, for those AUs with very low occurrence rates, directly predicting them as nonoccurrences results in poor individual

Table 7.4 – F1 score result comparison with state-of-the-art methods on DISFA dataset

Reference	**FACS - AUS**										Avg.
	1	**2**	**4**	**9**	**12**	**15**	**17**	**20**	**25**	**26**	
AlexNet	0.12	0.12	0.28	0.12	0.30	-	-	-	0.44	0.28	0.24
EAC-NET	0.410	0.264	0.664	0.805	0.893	-	-	-	0.889	0.156	0.485
Han et al. (2018)	0.437	0.400	0.672	0.497	0.758	0.378	0.523	-	0.724	0.548	0.55
Li et al. (2017)	0.426	0.272	0.655	0.228	0.829	-	-	-	0.883	0.259	0.51
Mei et al. (2018)	0.525	0.681	0.475	0.493	0.738	-	-	-	0.850	0.751	0.65
EvoNet	0.21	0.21	0.39	0.14	0.72	-	-	-	0.48	0.32	0.35
ARL	0.439	0.421	0.636	0.400	0.726	-	-	-	0.952	0.668	0.587
STRAL	0.522	0.474	0.689	0.567	0.725	-	-	-	0.913	0.676	0.63
Sankaran et al. (2020)	0.448	0.417	0.529	0.507	0.724	0.377	0.458	-	0.822	0.609	0.491
Ours CNN based	0.1	0.29	0.74	0.36	0.69	0.22	0.43	0.63	0.71	0.56	0.50
Our SqueezeNet based	0.31	0.70	0.40	0.42	0.70	0.40	0.50	0.5	0.76	0.6	0.529

accuracy but high average F1-frame results. Moreover, there is a more serious data imbalance issue in DISFA database, which results in large performance fluctuations over different AUs for most of the previous approaches (SHAO *et al.*, 2020a), including ours.

7.2 Discussion of Results

From the first experiment, we observed that applying geometric features into region learning is more important for facial expression classification because they contain more expression-selective information. We also find that selection based on face region is more effective than classification based on the whole face. This conclusion is in agreement with the region's learning theory underlying FACS AU recognition. It suggests that humans and CNNs use similar visual cues for facial expression recognition.

In our approach, we consider the facial occlusion in sign language by annotation, i.e., by the usage of FACS visibility codes (see Figure 7.3). Thereby, when the algorithm cannot detect the face, the output is the visibility code AU74, which means unscorable. However, if the occlusion is due to the hand being in front of the face, and consequentially, occluding the facial expression, the framework output is AU73, which translates to the entire face not visible (SILVA *et al.*, 2020a).

Regardless of the overall improvement of SqueezeNet across the 119 facial expressions, in Table 7.3, we notice that AU35 and AU33 had a lower classification rate. That can be justified by the fact that the facial action of AU18 is very similar to AU33 and AU35, as shown in Figure 7.3.

Still in Table 7.3, we notice that AU62 receives a very low classification rate, and that is due to the occurrence of AU62 being associated with the presence of AU51 and AU52. In Figure 7.4, we present examples of that association. Such combination with

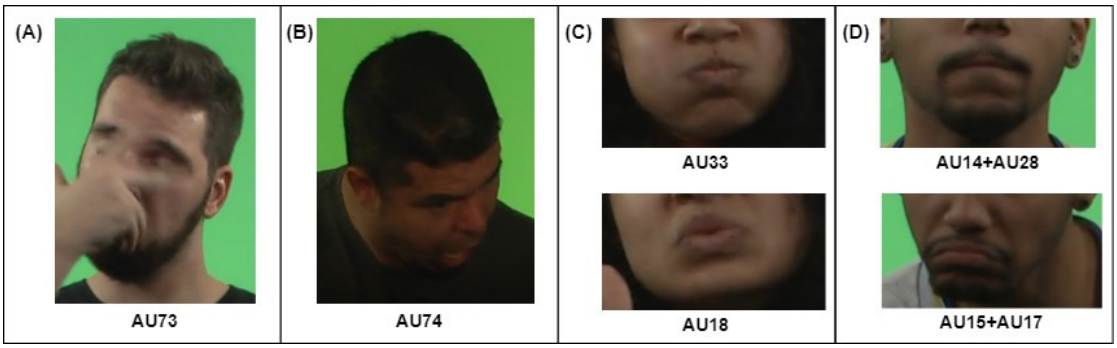


Figure 7.3 – Images (A) and (B) show examples of visibility codes. In images (C) and (D), we present samples of AU18, AU33, AU14+AU28, and AU15+AU17 where the Libras-SqueezeNet model showed confuse output. Images extracted from (SILVA *et al.*, 2020b).

Figure 7.4 – Example of AU51+AU61 from SILFA dataset (SILVA *et al.*, 2020b).

head movement makes difficult for our approach discerns from the other eye position. A similar situation occurs with AU19, which is the apparent tongue descriptor. In this case, AU19 is associated with the presence of AU25, in the sense that AU19 cannot occur without AU25, making difficult recognition of by itself.

Lastly, it also needs to account for the edge frames. It is called edge frames or fringe frames when a facial expression merges into the next, and it is not possible to differentiate the facial image portrayed in the frame. That occurred on 2031 frames on our data set, resulting in images poorly classified, augmenting the error rate. In most cases, they receive the annotation AU0 while their true label is AU73 or AU74.

Summarizing, the results described in Experiments 1–3 support our proposed combination of frameworks for Libras FACS classification. Particularly, due to the positive results obtained with the SqueezeNet based framework, we chose that to be the backbone structure of our method, and call it SqueezeNet-Libras. Figure 7.5 brings the full scheme for the SqueezeNet-Libras. Each frame from the input video is feed into the preprocessing stage where the face is detected and cropped into upper and lower parts. The geometric face features extracted are added as grayscale to the cropped image and further passed into a SqueezeNet model accordingly with the upper and lower part from the face. The output of those networks are combined into a label and converted into a FACS tier for the video input.

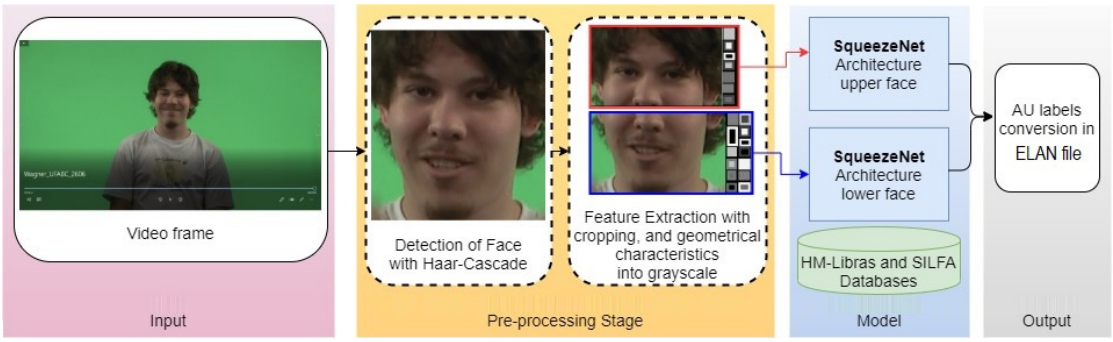


Figure 7.5 – Framework for SqueezeNet-Libras. Image created for the study itself.

7.3 Libras' Facial Action Annotation Analysis

So far, we explore the design and definition of our backbone network. Now, we analyzed the exit annotations. It is intuitive to us that even though the algorithm achieved a high hit rate in the datasets, that it may not achieve something similar in the annotation of the videos. To implement our annotator, with our SqueezeNet framework trained, the outputs are converted into a text file that can be read by ELAN. We can see the framework of our complete method in the Figure 7.6. With the defined structure, our system was tested on five videos that were recorded from the Sign Language Facial Action (SILFA) corpus (SILVA *et al.*, 2020b). These videos were not previously annotated. Two analysis were made: (1) The automatic AU annotation output was just later verified by a FACS trained annotator, and (2) the Libras' facial expressions function was directly inferred from the Libras-SqueezeNet AU output to be compared to two trained Libras' annotators.

In order to set the experiment, we need a measure of the inter-annotator agree-

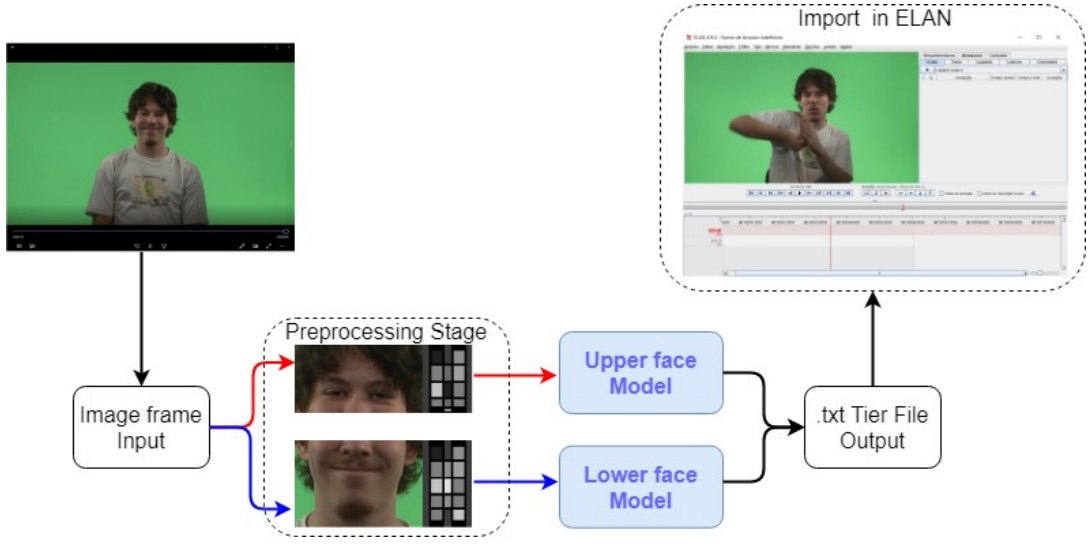


Figure 7.6 – The complete diagram of the FACS annotation method in Libras. Image created for the research itself.

Table 7.5 – Average percentage and frame count comparison between human annotator and our recognition output for each one of the five test videos in the Silfa corpus.

	# of frames	# of frames that both annotators agreed upon labeling	# of frames with correct labels	# of frames that needed correction	# of frames with face occlusion	Avg %
Video 1	240	226	136	100	6	0.57
Video 2	210	188	130	78	4	0.62
Video 3	210	182	115	91	4	0.55
Video 4	180	148	124	56	-	0.69
Video 5	210	180	140	70	-	0.67
Total	1050	924	645	395	14	0.62

ment[1]. A labeling agreement measures the similarity between annotations from different sources. To calculate the labeling agreement, we use percentages along with sets based on the number of frames for each video, as presented in Table 7.5. There were 1050 frames in the experimental set of unannotated videos. In $1,4\%$ of the frames, it was not possible the detection of the face due to occlusion, and they were annotated with AU73, in accordance with the human FACS coder. The total labeling agreement rate of the frames is 62%, obtained by comparing the labeling between the output of our methodology and the human coder. Table 7.5 also brings, as columns, the numbers of frames that both annotators agreed-upon labeling, the numbers of frames with correct labels (i.e., the number of frames that had corrected tags from the automatic tool accordingly with both human coders), the numbers of frames that needed correction (i.e., the number of frames that the automatic tool got the labels wrong, accordingly with both human coders), and the numbers of frames where face occlusion occurs for each video. The first advantage of this protocol is the improvement in reliability and in reproducibility. As the results show, the analysis from our annotations outputs compared to the one human labeled reveals relatively confidence agreement towards the automated tool.

For an enlightened comparison, the evaluation performed on five videos from our training SILFA dataset reveals an observed agreement of 82% between the fully manual annotations and the automatic SqueezeNet-Libras outputs. The other advantage of this automatic methodology is the gain in annotation time. The coder took a whole day of work to annotate and verify the videos, while our methods took around twenty-five minutes. This time spent manually correcting the automatic output takes less time than when our procedure was fully manual. The improvement of Libras' FACS annotation reliability and the time saved are the two main methodological contributions of this study.

[1] Inter-Annotator Agreement (IAA) measure of how well multiple annotators can make the same annotation decision for a certain category that have subjective interpretations(BRUIJN, 2018). Particularly, linguistic categories are determined by human judgment, making it challenging to measure correctness directly. Between annotators, a measure of reliability is done using coefficients of agreement, for instance, simple percentage, or Cohen and Fleiss' kappa (VIERA *et al.*, 2005).

This will allow the scientific community to explore larger corpora since it is often the annotation cost that may restraint a broader analysis.

To measure the Libras' facial expressions function classes inferred from our SqueezeNet-Libras output (see Figure 7.7) against the human annotation, we propose to feed fifteen videos from the SILFA corpus (that was not used in the training phase of our network), containing each type of Libras' facial expression function classes. Table 7.6 brings the percentage scores that were collected by comparing the annotation between the human annotators and the Libras-SqueezeNet output. The second column of Table 7.6 presents the percentage score, and it specifies the employment of the non-manuals markers, i.e., the number of times that an AU was corrected labeled to mark a grammatical construction. When analyzing the results encountered, we obtained above 60% of average accuracy. The numbers in the last column specify how many of our sentences can be correctly classified using combined AU recognition, which may refer to how descriptive the classification of the feature is. Some discriminant characteristics will, of course, be more common and, hence, will successfully classify more samples of facial expressions than others. For example, "open mouth" is a usual AU of GFE of norm (since 42% of our GFE of norm sentences are successfully classified with it), but is also employed elsewhere (if one were to use only this feature, then 85% of other sentences would also be classified as GFE of norm). In comparison, "one eye closed" is not a common AU for GFE of norm (only used in 13% of the sentences), but it is rarely used elsewhere (2% of other sentences). So it makes a very efficient and robust stand-alone non-manual marker to indicate that a sign belongs to the grammatical expression of norm category (with classification accuracy at 71%). Thus, the first non-manual marker is not as robust and descriptive as the last one.

Moreover, note in Table 7.6 that the labeling percentage for grammatical facial expression of intensity and grammatical facial expression of sentence for interrogative (WH-question, YN-question, and doubt) had the worst values. That can be assigned to the discrepancy occurred in transition frames or border errors, which are classification errors that happen inside a transition band between the occurrence of an onset, apex, and offset phase of facial expression, or even when there is rapid head movement involved in the performance of the sign. We also observed some output confusion in specific expressions in the grammatical facial expression of intensity regarding puffed cheeks and projected lips, which may negatively influence the results. Particularly in the GFE of intensity case, our analysis showed a rather considerable confusion on classifying the "intensity face". Examples of such expressions are shown in Fig 7.3. The last rows of Table 7.6 present the occlusion rates as almost fair confidence for our framework.

Finally, the proposed Libras-SquezeNet AU classification framework, which

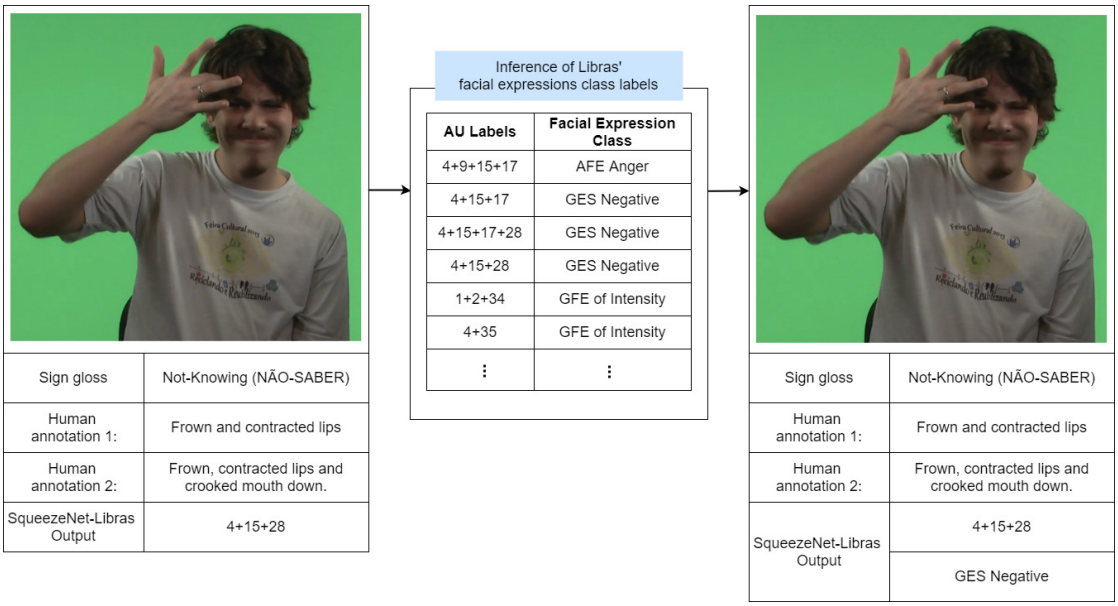


Figure 7.7 – The illustration shows a frame sample from the SILFA dataset annotated by two human coders and the exit from SqueezeNet-Libras. Based on the output, our system estimate that the facial expression belongs to the GFE of Sentence Negative class, creating a new labeling tier. Image created for the research itself.

Table 7.6 – Results obtained with the Libras-SqueezeNet model for each of the Libras' Non-Manual Markers

Facial Expression Type	**Avg. Acc**
GES WH-Question	0.62
GES YN-Question	0.63
GES Doubt	0.65
GES Topic	0.70
GES Negative	0.72
GES Assertive	0.70
GES Focus	0.69
GES Condicional Clause	0.69
GES Relative Clause	0.68
Grammatical expression of intensity	0.66
Grammatical expression of Homonymy	0.73
Grammatical expression of norm	0.71
Affective Facial Expression	0.76
Neutral	0.95
Face Occlusion	0.67
Unscorable	0.80

uses geometric and region of interest features, implements the first automatic annotator of Libras' Non-Manual Markers. Yet, when averaging our prediction with the manual annotations, the performance can be further improved. Combined with the previously presented experimental results, we can imply that learning Libras' facial expressions as a function of compound AUs may be a systematic and attractive alternative way than learning facial expressions from the whole face.

7.4 Failure analysis

Before we conclude our work, a description of the uneventful or intercurrent events in this research is required, not only as a precautionary tale but also as a more accurate idea of our research path.

In the modeling stage, we first tried to implement the facial expression labeling described in McCleary (2010)(MCCLEARY *et al.*, 2010) and Freitas *et al.* (2011)(FREITAS, 2011). However, the symbols used in both approaches introduced implementation errors when importing from ELAN annotations to the template in python. Another coding for facial expressions in Libras is described in Domingues (2015)(DOMINGUES, 2015), where they bring facial expression terminologies into three categories: sensations and positive feelings (SSP), sensations and negative feelings (SSN), and states and face movements (EMF). Yet, their coding does not cover all the facial behavior we had previously found in the literature survey, making necessary a large creation of new codes.

Another attempt made was to test with different networks structure. In the CNN case, we trial different combinations of the number of layers and neurons in the fully connected layer. The previously presented CNN configuration obtained the best results and was the chosen one to be detailed in the text. Also, we implemented MobileNetV1(HOWARD *et al.*, 2017) and ResNet-50(HE *et al.*, 2016) architectures. Both networks also were trained from scratch, with fine-tuned parameters, and they performed worst than the networks presented. When put in a larger perspective, these architectures are fairly deep in contrast to the ones earlier presented, so this light analysis of our ill-fitting approaches can be informative for future works to avoid.

7.5 Concluding Remarks

In conclusion, the results of conducted experiments have been evaluated and analysed. SqueezeNet-Libras outperformed all other methods with its a more exceptional ability to learn facial features. There was a definite relationship observed between the preprocessing stage and the accuracy of the classification task with a CNN. The chosen preprocess stage provided to train shown a better performance of the adopted AUs detection and recognition task on the SqueezeNet network. We also compared our results with the current state-of-art work in the DISFA database.

Conclusion

Aimed to contribute to the development of Automatic Sign Language Recognition (ASLR) technologies, the present work proposed a novel facial expression recognition model trained with Brazilian sign language facial expressions. Besides, our possible applications comprehend the fields of education, health care, marketing, security, and many others where analysis and classification of facial expressions have applications.

The challenges faced by the work included the lack of a well established consensus regarding the identification, the description, and the function of facial expressions in Libras. To overcome this obstacle, we conducted a comprehensive review of the documented facial expressions in Libras, we proposed a taxonomy for the grammatical functions played by the facial expressions in Libras, and we encoded them using Facial Action Coding System (FACS) (Chapters 3 and 4).

Following, we approached the problem of building annotated databases of sign language facial expressions. As described in Chapter 5, we built HM-Libras and SILFA, two novel datasets annotated with FACS. SILFA is a dataset constructed under controlled conditions, while HM-Libras is a collection of videos extracted from the Internet.

To identify the best network architecture for the classification of Libras' facial action units, we performed tests with various models (Chapters 6 and 7). To evaluate their efficacy we compared the results of our prototype models with well established trained networks, most of them designed to recognize facial expressions of emotions. Our experiments showed, for example, that a preprocessing stage was capable of helping the neural network to to learn better characteristic descriptors. Also, in the study of simple facial action units, the usage of compound facial expressions for training, in contrast with individual action units, stood out in relation to the results found in state of the art by the same experimental setup. Such findings verify our hypothesis that the compound facial expression would improve simple facial expression recognition, something that was just hypothesized before our work. Despite this verification, a great part of our average result in the f-score was not above state of the art. That could be traced to the fact that our framework got confused predictability between facial expressions that are too similar. Also, the nature of head movement being closed to eye gaze direction in sign language proved a challenge for our FER recognition.

We also developed a software application that translates the output of our Facial Expression Recognition (FER) model into ELAN (Eudico Linguistic Annotator) files. This automatic annotation process can be used with images and videos of any sign

language. In Chapter 7 we compared the output of our system with the labels given by human annotators. We realized that our model still misses the classification of facial expressions that are very similar when there are pose variation or other action units present in the face. Still, our experiments achieved a 62% labeling agreement rate, demonstrating that the novel methodology solution for the presence of 119 action units is impressive, feasible, and can be improved.

In summary, the principal contributions of this work are:

- The proposal of a novel AU classifier and the first automatic annotation system for Libras' facial expressions, including their grammatical function;
- The presentation of a taxonomy for facial expressions in Libras, derived from a comprehensive review of the literature, and their encoding using FACS;
- The building of two annotated datasets of facial expressions in Libras;

Opportunities ahead and next steps include:

- The proposed framework does not take into account the temporal dimension of the input data (frames of videos). Assuming time-domain information could improve the results, a recurrent network could readily fit into our framework.
- More extensive experimentation with alternative preprocessing techniques could be carried out. Precisely, to extract different types of information regarding the head position to improve tracking and pose recognition.
- Another limiting factor was the opportunity of using the relation between the simple and compound facial expression. In Sign Language, is possible to infer such relation from context. So the proposal of attention layers into the architecture could bring better results. The aim is to prevent similar facial expressions from confusing the output prediction.
- Further exploration by comparative analysis with our SILFA data set, testing in other sign languages (e.g., ASL), checking for differences in non-manual markers, and if our proposal behaves well in other sign languages.

We hope that this type of multidisciplinary study become a sparkle that could ignite other studies and other applications. We believe that our work could help to improve the speed and the descriptive capacity of video annotation that may leverage objective sign language linguistic studies. Also, associated with automatic gesture recognition models, our work is a further step in the pursuit of a complete solution to the ASLR problem,

which will certainly contribute to making our Brazilian deaf community more included in the hearing society and vice-versa.

Bibliography

AGIANPUYE, S.; MINOI, J.-L. 3d facial expression synthesis: A survey. In: IEEE. *2013 8th International Conference on Information Technology in Asia (CITA)*. [S.l.], 2013. p. 1–7. Cited on page 21.

AGRIS, U. V.; KNORR, M.; KRAISS, K.-F. The significance of facial features for automatic sign language recognition. In: IEEE. *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. [S.l.], 2008. p. 1–6. Cited 2 times on pages 50 and 64.

AHMED, A. M.; ALEZ, R. A.; THARWAT, G.; TAHA, M.; GHRIBI, W.; BADAWY, A. S.; CHANGALASETTY, S. B.; BOSE, J. S. C. Towards the design of automatic translation system from arabic sign language to arabic text. In: IEEE. *Inventive Computing and Informatics (ICICI), International Conference on*. [S.l.], 2017. p. 325–330. Cited on page 69.

AIFANTI, N.; PAPACHRISTOU, C.; DELOPOULOS, A. The mug facial expression database. In: IEEE. *Image analysis for multimedia interactive services (WIAMIS), 2010 11th international workshop on*. [S.l.], 2010. p. 1–4. Cited on page 44.

ALBRES, N. A.; COSTA, M. P. P.; ROSSI, T. W. T. Gesto-visualidade no processo de tradução de literatura infanto-juvenil: marcas do discurso narrativo. *Translatio*, n. 9, p. 3–20, 2015. Cited on page 58.

ALMEIDA, S. G. M. Extração de características em reconhecimento de parâmetros fonológicos da língua brasileira de sinais utilizando sensores rgb-d. Universidade Federal de Minas Gerais, 2014. Cited 3 times on pages 54, 57, and 137.

ALSTON, W. P.; ALSTON, Y. *Illocutionary acts and sentence meaning*. [S.l.]: Cornell University Press, 2000. Cited on page 72.

ANTONAKOS, E.; ROUSSOS, A.; ZAFEIRIOU, S. A survey on mouth modeling and analysis for sign language recognition. In: IEEE. *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. [S.l.], 2015. v. 1, p. 1–7. Cited 2 times on pages 51 and 64.

ARAN, O.; BURGER, T.; CAPLIER, A.; AKARUN, L. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition*, Elsevier, v. 42, n. 5, p. 812–822, 2009. Cited on page 50.

ARAUJO, A. D. d. *As expressões e as marcas não-manuais na Língua de Sinais Brasileira*. Dissertação (Mestrado) — Universidade de Brasília (UnB), 2013. Cited 7 times on pages 56, 59, 60, 61, 64, 65, and 137.

ARROTÉIA, J. *O papel da marcação não-manual nas sentenças negativas em Língua de Sinais Brasileira (LSB)*. Dissertação (Mestrado) — Universidade Estadual de Campinas, Novembro 2005. Cited 6 times on pages 54, 56, 57, 59, 61, and 137.

ASSOCIATION, N. C. S. *et al. Cued Speech and Literacy: History, Research, and Background Information*. [S.l.], 2006. Cited on page 51.

BA, S. O.; ODOBEZ, J. M. Visual focus of attention estimation from head pose posterior probability distributions. In: IEEE. *Multimedia and Expo, 2008 IEEE International Conference on*. [S.l.], 2008. p. 53–56. Cited on page 52.

BAILLY, K.; MILGRAM, M. Bisar: Boosted input selection algorithm for regression. In: IEEE. *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. [S.l.], 2009. p. 249–255. Cited on page 52.

BAKER, C.; COKELY, D. Asl: A teacher's resource text on grammar and culture. *Silver Spring, MD: TJ Publishers*, 1980. Cited on page 54.

BAKER, C.; PADDEN, C. *Focusing on the nonmanual components of American Sign Language. Understanding language through sign language research, ed. by P. Siple, 27-57*. [S.l.]: New York: Academic Press, 1978. Cited on page 58.

BAKER-SHENK, C. A microanalysis of the nonmanual components of questions in american sign language. 1983. Cited on page 56.

BAKER-SHENK, C. The facial behavior of deaf signers: Evidence of a complex language. *American Annals of the Deaf*, Gallaudet University Press, v. 130, n. 4, p. 297–304, 1985. Cited on page 22.

BALNTAS, V.; RIBA, E.; PONSA, D.; MIKOLAJCZYK, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In: *Bmvc*. [S.l.: s.n.], 2016. v. 1, n. 2, p. 3. Cited 2 times on pages 40 and 49.

BANK, R.; CRASBORN, O.; HOUT, R. V. Alignment of two languages: The spreading of mouthings in sign language of the netherlands. *International Journal of Bilingualism*, Sage Publications Sage UK: London, England, v. 19, n. 1, p. 40–55, 2015. Cited on page 51.

BANK, R.; CRASBORN, O. A.; HOUT, R. V. Variation in mouth actions with manual signs in sign language of the netherlands (ngt). *Sign Language & Linguistics*, John Benjamins, v. 14, n. 2, p. 248–270, 2011. Cited on page 51.

BARROS, M. E. Princípios básicos da elis: escrita das línguas de sinais. Sueli Maria Regino, 2016. Cited on page 70.

BARROS, M. E. *et al.* Elis-escrita das línguas de sinais: proposta teórica e verificação prática. Florianópolis, SC, 2008. Cited on page 70.

BARTLETT, M. S.; LITTLEWORT, G.; LAINSCSEK, C.; FASEL, I.; MOVELLAN, J. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: IEEE. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*. [S.l.], 2004. v. 1, p. 592–597. Cited 2 times on pages 41 and 42.

BATISTA, J. C.; ALBIERO, V.; BELLON, O. R.; SILVA, L. Aumpnet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. In: IEEE. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. [S.l.], 2017. p. 866–871. Cited on page 46.

BATTISON, R. Phonological deletion in american sign language. *Sign language studies*, Gallaudet University Press, v. 5, n. 1, p. 1–19, 1974. Cited on page 58.

BENITEZ-QUIROZ, C. F.; GÖKGÖZ, K.; WILBUR, R. B.; MARTINEZ, A. M. Discriminant features and temporal structure of nonmanuals in american sign language. *PloS one*, Public Library of Science, v. 9, n. 2, p. e86268, 2014. Cited on page 51.

BENITEZ-QUIROZ, C. F.; SRINIVASAN, R.; FENG, Q.; WANG, Y.; MARTINEZ, A. M. *EmotioNet Challenge: Recognition of facial expressions of emotion in the wild*. 2017. Cited 2 times on pages 46 and 49.

BÉRARD, A.; BESACIER, L.; KOCABIYIKOGLU, A. C.; PIETQUIN, O. End-to-end automatic speech translation of audiobooks. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2018. p. 6224–6228. Cited on page 23.

BEREZ, A. L. Review of eudico linguistic annotator (elan). University of Hawai'i Press, 2007. Cited 2 times on pages 75 and 84.

BERNARDINO, E. L. A. O uso de classificadores na língua de sinais brasileira. *ReVEL*, v. 10, n. 19, 2012. Cited on page 58.

BHUVAN, M. S.; RAO, D. V.; JAIN, S.; ASHWIN, T.; GUDDETTI, R. M. R.; KULGOD, S. P. Detection and analysis model for grammatical facial expressions in sign language. In: IEEE. *Region 10 symposium (tensymp), 2016 ieee*. [S.l.], 2016. p. 155–160. Cited 2 times on pages 24 and 52.

BISHOP, C. M. *et al. Neural networks for pattern recognition*. [S.l.]: Oxford university press, 1995. Cited 2 times on pages 31 and 33.

BOIS, J. W. D. Discourse transcription. *Santa Barbara papers in linguistics*, v. 4, p. 1–225, 1992. Cited on page 70.

BORYSOVA, A. Leap motion controller for sign language recognition: A review of the literature. 2017. Cited on page 51.

BOTTOU, L. Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010*. [S.l.]: Springer, 2010. p. 177–186. Cited on page 37.

BRADSKI, G.; KAEHLER, A. Opencv. *Dr. Dobb's journal of software tools*, v. 3, 2000. Cited 3 times on pages 12, 89, and 90.

BRASIL, A. Libras: dicionário da língua basileira de sinais. *Acessibilidade Brasil. Software Versão*, v. 2, 2006. Cited on page 83.

BRAZIL. *Decree-Law No.10,436, of April 24, 2002.* 2002. <http://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm>. Cited on page 54.

BRUIJN, L. de. *Inter-Annotator Agreement (IAA)*. 2018. <https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3>. Cited on page 107.

CAPOVILLA, F. C. Visemas e quiremas e outros bípedes implumes: revisão etimológica da taxonomia da linguagem em surdez - por que e como fazer. Educação digital: a tecnologia a favor da inclusão. Porto Alegre, Penso, p. 239–262, 2013. Cited on page 58.

CAPOVILLA, F. C.; GARCIA, W. Visemas, quiremas, e bípedes implumes: Por uma revisão taxonômica da linguagem do surdo que substitua visemas por fanerolaliemas, e quiremas por simatosemas para forma de mão (quiriformemas), local de mão (quiritoposema), movimento de mão (quiricinesema), e expressão facial (mascarema). *Transtornos de aprendizagem*, v. 2, p. 96–101, 2011. Cited on page 60.

CAPOVILLA, F. C.; GARCIA, W. O. Análise da estrutura sematosêmica-signumicular do corpus de 10.338 sinais da 3ª ed. do novo deit libras via buscasigno, versão 2. *Novo Deit-Libras: Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira (Libras) baseado em Linguística e Neurociências Cognitivas*, v. 2, p. 2684–2701, 2012. Cited on page 60.

CAPOVILLA, F. C.; RAPHAEL, W. D.; MAURICIO, A. C. *Novo dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira (Novo Deit-Libras)*. [S.l.]: Edusp, 2008. Cited 3 times on pages 64, 65, and 66.

CAPOVILLA, F. C.; RAPHAEL, W. D.; TEMOTEO, J. G.; MARTINS, A. C. Dicionário da língua de sinais do brasil: A libras em suas mãos. *São Paulo: Editora da Universidade de São Paulo*, v. 1, 2017. Cited 8 times on pages 55, 64, 65, 66, 73, 83, 86, and 137.

CARDOSO, M. E. d. A. *Segmentação automática de Expressões Faciais Gramaticais com Multilayer Perceptrons e Misturas de Especialistas*. Dissertação (Mestrado) — Universidade de São Paulo, 2018. Cited 2 times on pages 21 and 22.

CARIDAKIS, G.; ASTERIADIS, S.; KARPOUZIS, K. Non-manual cues in automatic sign language recognition. *Personal and ubiquitous computing*, Springer, v. 18, n. 1, p. 37–46, 2014. Cited on page 50.

CARNEIRO, B. G. O corpo na concepção de eventos na língua de sinais brasileira. *ANTARES: Letras e Humanidades*, v. 7, n. 14, 2015. Cited on page 58.

CHAN, H.; BLEDSOE, W. A man-machine facial recognition system: some preliminary results. *Panoramic Research Inc., Palo Alto, CA, USA1965*, 1965. Cited on page 29.

CHENGETA, K.; VIRIRI, S. A survey on facial recognition based on local directional and local binary patterns. In: IEEE. *2018 Conference on Information Communications Technology and Society (ICTAS)*. [S.l.], 2018. p. 1–6. Cited on page 40.

CHIBELUSHI, C. C.; BOUREL, F. Facial expression recognition: A brief tutorial overview. *CVonline: On-Line Compendium of Computer Vision*, v. 9, 2003. Cited on page 29.

CHOLLET, F. *et al.* Keras: The python deep learning library. *Astrophysics Source Code Library*, 2018. Cited on page 98.

CHU, W.-S.; TORRE, F. De la; COHN, J. F. Learning spatial and temporal cues for multi-label facial action unit detection. In: IEEE. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. [S.l.], 2017. p. 25–32. Cited 4 times on pages 42, 47, 48, and 97.

CHU, W.-S.; TORRE, F. De la; COHN, J. F. Learning facial action units with spatiotemporal cues and multi-label sampling. *Image and Vision Computing*, Elsevier, v. 81, p. 1–14, 2019. Cited 4 times on pages 12, 42, 93, and 94.

COHEN, I. *Automatic facial expression recognition from video sequences using temporal information*. Dissertação (Mestrado) — University of Illinois at Urbana-Champaign, 2000. Cited on page 41.

CONTI-RAMSDEN, G. Clan (computerized language analysis). *Child Language Teaching and Therapy*, Sage Publications Sage CA: Thousand Oaks, CA, v. 12, n. 3, p. 345–349, 1996. Cited on page 74.

COOK, K. A.; THOMAS, J. J. *Illuminating the path: The research and development agenda for visual analytics*. [S.l.], 2005. Cited on page 59.

COOPER, H.; HOLT, B.; BOWDEN, R. Sign language recognition. In: *Visual Analysis of Humans*. [S.l.]: Springer, 2011. p. 539–562. Cited 2 times on pages 50 and 51.

CRASBORN, O. A. Nonmanual structures in sign language. *Encyclopedia of Language Linguistics*, Oxford: Elsevier, v. 8, p. 668–672, 2006. Cited on page 64.

CROWTHER, J. *Oxford advanced learner's dictionary of current English*. [S.l.]: Oxford University Press, 1995. Cited on page 69.

DACHKOVSKY, S.; SANDLER, W. Visual intonation in the prosody of a sign language. *Language and speech*, SAGE Publications Sage UK: London, England, v. 52, n. 2-3, p. 287–314, 2009. Cited on page 21.

DARWIN, C. *The expression of the emotions in man and animals*. [S.l.]: John Murray, United Kingdom, 1872. Cited on page 29.

DENG, D.; CHEN, Z.; SHI, B. E. Fau, facial expressions, valence and arousal: A multi-task solution. *arXiv preprint arXiv:2002.03557*, 2020. Cited 2 times on pages 42 and 49.

DERCZYNSKI, L. Complementarity, f-score, and nlp evaluation. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. [S.l.: s.n.], 2016. p. 261–266. Cited on page 49.

DESHPANDE, A. *A Beginner's Guide To Understanding Convolutional Neural Networks*. 2016. <https://adeshpande3.github.io/>. Cited on page 94.

DOMINGUES, K. N. P. *Análise de estrutura SematosÊmica de 10.400 sinais de Libras: caracterização das combinações canônicas entre articulação de mão, orientações de mão e palma, movimento, e expressão facial*. Tese (Doutorado) — Universidade de São Paulo, 2015. Cited 2 times on pages 60 and 110.

DONAHUE, J.; HENDRICKS, L. A.; GUADARRAMA, S.; ROHRBACH, M.; VENUGOPALAN, S.; SAENKO, K.; DARRELL, T. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 2625–2634. Cited on page 48.

DREWES, H. *Eye gaze tracking for human computer interaction*. Tese (Doutorado) — lmu, 2010. Cited on page 51.

DREWES, H.; SCHMIDT, A. Interacting with the computer using gaze gestures. In: SPRINGER. *IFIP Conference on Human-Computer Interaction*. [S.l.], 2007. p. 475–488. Cited on page 51.

DU, S.; TAO, Y.; MARTINEZ, A. M. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 111, n. 15, p. E1454–E1462, 2014. Cited 3 times on pages 11, 21, and 43.

EIA. *Emotional Intelligence Academy*. 2019. <https://www.eiagroup.com/courses/>. Cited on page 75.

EKMAN, P. Are there basic emotions? American Psychological Association, 1992. Cited on page 44.

EKMAN, P. Facial expression and emotion. *American psychologist*, American Psychological Association, v. 48, n. 4, p. 384, 1993. Cited 2 times on pages 21 and 44.

EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, American Psychological Association, v. 17, n. 2, p. 124, 1971. Cited 2 times on pages 29 and 41.

EKMAN, P.; FRIESEN, W. V. *Manual for the facial action coding system*. [S.l.]: Consulting Psychologists Press, 1978. Cited 5 times on pages 21, 41, 43, 75, and 76.

EKMAN, P.; FRIESEN, W. V. *Unmasking the face: A guide to recognizing emotions from facial clues*. [S.l.]: Ishk, 2003. Cited on page 21.

FAN, H.; JIANG, M.; XU, L.; ZHU, H.; CHENG, J.; JIANG, J. Comparison of long short term memory networks and the hydrological model in runoff simulation. *Water*, Multidisciplinary Digital Publishing Institute, v. 12, n. 1, p. 175, 2020. Cited 2 times on pages 12 and 94.

FASEL, B.; LUETTIN, J. Automatic facial expression analysis: a survey. *Pattern recognition*, Elsevier, v. 36, n. 1, p. 259–275, 2003. Cited 2 times on pages 38 and 39.

FELIPE, T. A. Introdução à gramática da libras. *Série Atualidades Pedagógicas*, v. 4, n. 3, p. 81–107, 1997. Cited 4 times on pages 51, 58, 70, and 71.

FELIPE, T. A. The verbalvisual discourse in brazilian sign language–libras. *Bakhtiniana: Revista de Estudos do Discurso*, v. 8, n. 2, 2013. Cited 3 times on pages 56, 64, and 137.

FERNANDEZ, P. D. M.; REN, T. I.; JYH, T. I.; PEÑA, F. A. G.; CUNHA, A. Deep metric structured learning for facial expression recognition. *arXiv preprint arXiv:2001.06612*, 2020. Cited on page 49.

FERREIRA-BRITO, L. Comparação de aspectos lingüísticos da lscb e do português. In: *Conferência apresentada no II Encontro Nacional de Pais e Amigos de Surdos. Porto Alegre*. [S.l.: s.n.], 1986. v. 27. Cited on page 23.

FERREIRA-BRITO, L. Integração social do surdo. *Trabalhos em Linguística Aplicada*, v. 7, 1986. Cited on page 54.

FERREIRA-BRITO, L. Uma abordagem fonológica dos sinais da lscb. *Informativo Técnico-Científico do INES, Rio de Janeiro*, v. 1, n. 1, p. 20–43, 1990. Cited 2 times on pages 54 and 56.

FERREIRA-BRITO, L. *Por uma gramática de línguas de sinais*. [S.l.]: Tempo Brasileiro, 1995. Cited 5 times on pages 54, 56, 61, 71, and 72.

FREITAS, F. A.; PERES, S. M.; LIMA, C. A. de M.; BARBOSA, F. V. Grammatical facial expressions recognition with machine learning. In: *FLAIRS Conference*. [S.l.: s.n.], 2014. Cited 11 times on pages 22, 24, 56, 59, 61, 64, 65, 66, 80, 81, and 137.

FREITAS, F. d. A. *Reconhecimento automático de expressões faciais gramaticais na língua brasileira de sinais*. Dissertação (Mestrado) — Universidade de São Paulo (USP), 2011. Cited 5 times on pages 24, 52, 59, 110, and 137.

FREITAS-MAGALHÃES, A. *Facial expression of emotion: from theory to application*. [S.l.]: Leya, 2013. Cited on page 21.

FRYDRYCH, L. A. K. Transcrição da interpretação para libras: uma abordagem enunciativa. 2010. Cited on page 73.

GALE, T.; ELSEN, E.; HOOKER, S. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019. Cited on page 46.

GEORGHIADES, A.; BELHUMEUR, P.; KRIEGMAN, D. Yale face database. *Center for computational Vision and Control at Yale University, http://cvc. yale. edu/projects/yalefaces/yalefa*, v. 2, p. 6, 1997. Cited on page 44.

GHAYOUMI, M.; BANSAL, A. K. Unifying geometric features and facial action units for improved performance of facial expression analysis. *arXiv preprint arXiv:1606.00822*, 2016. Cited 2 times on pages 41 and 46.

GHOSH, S.; LAKSANA, E.; SCHERER, S.; MORENCY, L.-P. A multi-label convolutional neural network approach to cross-domain action unit detection. In: IEEE. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. [S.l.], 2015. p. 609–615. Cited on page 47.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. [S.l.: s.n.], 2011. p. 315–323. Cited on page 32.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Cited 2 times on pages 35 and 36.

GUDI, A.; TASLI, H. E.; UYL, T. M. D.; MAROULIS, A. Deep learning based facs action unit occurrence and intensity estimation. In: IEEE. *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. [S.l.], 2015. v. 6, p. 1–5. Cited 3 times on pages 42, 47, and 48.

HAMM, J.; KOHLER, C. G.; GUR, R. C.; VERMA, R. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, Elsevier, v. 200, n. 2, p. 237–256, 2011. Cited on page 47.

HAN, S.; MENG, Z.; LI, Z.; O'REILLY, J.; CAI, J.; WANG, X.; TONG, Y. Optimizing filter size in convolutional neural networks for facial action unit recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 5070–5078. Cited on page 103.

HANKE, T. Hamnosys-representing sign language data in language resources and language processing contexts. In: *LREC*. [S.l.: s.n.], 2004. v. 4, p. 1–6. Cited on page 70.

HAPPY, S.; PATNAIK, P.; ROUTRAY, A.; GUHA, R. The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*, IEEE, v. 8, n. 1, p. 131–142, 2017. Cited on page 44.

HARRIGAN, J.; ROSENTHAL, R.; SCHERER, K. R.; SCHERER, K. *New handbook of methods in nonverbal behavior research*. [S.l.]: Oxford University Press, 2008. Cited on page 43.

HAYKIN, S. S.; HAYKIN, S. S.; HAYKIN, S. S.; HAYKIN, S. S. *Neural networks and learning machines*. [S.l.]: Pearson Upper Saddle River, 2009. v. 3. Cited 6 times on pages 30, 31, 32, 33, 34, and 37.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Cited on page 110.

HERACLEOUS, P.; ABOUTABIT, N.; BEAUTEMPS, D. Lip shape and hand position fusion for automatic vowel recognition in cued speech for french. *IEEE Signal Processing Letters*, IEEE, v. 16, n. 5, p. 339–342, 2009. Cited on page 51.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Cited on page 34.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. Cited on page 110.

IANDOLA, F. N.; HAN, S.; MOSKEWICZ, M. W.; ASHRAF, K.; DALLY, W. J.; KEUTZER, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. Cited on page 95.

IBRAHIM, N. B.; ZAYED, H. H.; SELIM, M. M. Advances, challenges and opportunities in continuous sign language recognition. *Journal of Engineering and Applied Sciences*, v. 15, n. 5, p. 1205–1227, 2020. Cited on page 50.

IMOTION. *Facial Action Coding System (FACS) – A Visual Guidebook*. 2019. <http://https://imotions.com/blog/facial-action-coding-system/>. Cited on page 75.

JAISWAL, S.; VALSTAR, M. Deep learning the dynamic appearance and shape of facial action units. In: IEEE. *2016 IEEE winter conference on applications of computer vision (WACV)*. [S.l.], 2016. p. 1–8. Cited on page 48.

JØRGENSEN, J. Imperatives and logic. *Erkenntnis*, Springer, v. 7, n. 1, p. 288–296, 1937. Cited on page 72.

KACORRI, H. Tr-2015001: A survey and critique of facial expression synthesis in sign language animation. City University of New York (CUNY), 2015. Cited on page 21.

KANADE, T.; TIAN, Y.; COHN, J. F. Comprehensive database for facial expression analysis. In: IEEE. *fg*. [S.l.], 2000. p. 46. Cited on page 45.

KATSIOS, D. *CNN Architectures: SQUEEZENET. Machine Learning Tokyo*. 2019. <https://machinelearningtokyo.com/2020/04/11/cnn-architectures-squeezenet/>. Cited 2 times on pages 12 and 96.

KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1867–1874. Cited 4 times on pages 12, 39, 89, and 90.

KELLY, D.; MCDONALD, J.; MARKHAM, C. Recognition of spatiotemporal gestures in sign language using gesture threshold hmms. In: *Machine Learning for Vision-Based Motion Analysis*. [S.l.]: Springer, 2011. p. 307–348. Cited on page 50.

KING, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, v. 10, p. 1755–1758, 2009. Cited on page 82.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Cited on page 37.

KIPP, M. Anvil-a generic annotation tool for multimodal dialogue. In: *Seventh European Conference on Speech Communication and Technology*. [S.l.: s.n.], 2001. Cited on page 74.

KOELSTRA, S.; PANTIC, M.; PATRAS, I. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 32, n. 11, p. 1940–1954, 2010. Cited on page 97.

KOHONEN, T. *Self-organization and associative memory*. [S.l.]: Springer Science & Business Media, 2012. v. 8. Cited on page 31.

KOLLER, O.; CAMGOZ, C.; NEY, H.; BOWDEN, R. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, 2019. Cited on page 51.

KOLLER, O.; FORSTER, J.; NEY, H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, Elsevier, v. 141, p. 108–125, 2015. Cited on page 51.

KOLLER, O.; NEY, H.; BOWDEN, R. Read my lips: Continuous signer independent weakly supervised viseme recognition. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2014. p. 281–296. Cited on page 51.

KOLLIAS, D.; SCHULC, A.; HAJIYEV, E.; ZAFEIRIOU, S. Analysing affective behavior in the first abaw 2020 competition. *arXiv preprint arXiv:2001.11409*, 2020. Cited on page 49.

KOLLIAS, D.; ZAFEIRIOU, S. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. Cited on page 49.

KRINIDIS, M.; NIKOLAIDIS, N.; PITAS, I. 3-d head pose estimation in monocular video sequences using deformable surfaces and radial basis functions. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 19, n. 2, p. 261–272, 2009. Cited on page 52.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105. Cited on page 103.

KUHNKE, F.; RUMBERG, L.; OSTERMANN, J. Two-stream aural-visual affect analysis in the wild. *arXiv preprint arXiv:2002.03399*, 2020. Cited 2 times on pages 42 and 49.

KUMADA, K. M. O. Libras: Língua brasileira de sinais. *Londrina: Editora e Distribuidora Educacional SA*, 2016. Cited 3 times on pages 11, 62, and 71.

KUMADA, K. M. O.; SILVA, I. R.; MARTINO, J. M. D.; COSTA, P. D. P. *et al.* Desafios da aprendizagem de português/libras por meio da tradução de material didático com uso de avatares expressivos: Foco na criação de conceitos para o ensino de ciências. In: *I Congresso Internacional de Educação Especial e Inclusiva/XII Jornada de Educação Especial (18-20 Maio, Marília)*. [S.l.: s.n.], 2016. Cited 6 times on pages 11, 55, 62, 64, 65, and 66.

LAN, Y.; HARVEY, R.; THEOBALD, B.; ONG, E.-J.; BOWDEN, R. Comparing visual features for lipreading. In: *International Conference on Auditory-Visual Speech Processing 2009*. [S.l.: s.n.], 2009. p. 102–106. Cited on page 51.

LANGNER, O.; DOTSCH, R.; BIJLSTRA, G.; WIGBOLDUS, D. H.; HAWK, S. T.; KNIPPENBERG, A. V. Presentation and validation of the radboud faces database. *Cognition and emotion*, Taylor & Francis, v. 24, n. 8, p. 1377–1388, 2010. Cited on page 44.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015. Cited 3 times on pages 31, 32, and 33.

LECUN, Y.; BENGIO, Y. *et al.* Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, v. 3361, n. 10, p. 1995, 1995. Cited on page 32.

LECUN, Y. A.; BOTTOU, L.; ORR, G. B.; MÜLLER, K.-R. Efficient backprop. In: *Neural networks: Tricks of the trade*. [S.l.]: Springer, 2012. p. 9–48. Cited on page 37.

LEE, M.; PAVLOVIC, V.; PANTIC, M. *et al.* Fast and effective adaptation of facial action unit detection deep model. *arXiv preprint arXiv:1909.12158*, 2019. Cited 2 times on pages 47 and 49.

LEMOS, A. M. As estratégias de interpretação de unidades fraseológicas do português para libras em discursos de políticos. www. teses. ufc. br, 2012. Cited on page 75.

LI, J.; WANG, J. Z. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 25, n. 9, p. 1075–1088, 2003. Cited on page 69.

LI, S.; DENG, W. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, p. 1–1, 2020. Cited 7 times on pages 11, 29, 38, 39, 40, 41, and 46.

LI, W.; ABTAHI, F.; ZHU, Z. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 1841–1850. Cited 3 times on pages 48, 93, and 103.

LI, W.; ABTAHI, F.; ZHU, Z.; YIN, L. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. *arXiv preprint arXiv:1702.02925*, 2017. Cited on page 47.

LI, W.; ABTAHI, F.; ZHU, Z.; YIN, L. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 40, n. 11, p. 2583–2596, 2018. Cited 2 times on pages 97 and 103.

LI, Y.; ZENG, J.; SHAN, S.; CHEN, X. Self-supervised representation learning from videos for facial action unit detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 10924–10933. Cited on page 49.

LIDDELL, S. K. Nonmanual signals and relative clauses in american sign language. *Understanding language through sign language research*, Academic Press New York, p. 59–90, 1978. Cited on page 58.

LIDDELL, S. K. *American sign language syntax*. [S.l.]: Mouton De Gruyter, 1980. v. 52. Cited on page 64.

LIDDELL, S. K.; JOHNSON, R. E. American sign language: The phonological base. *Sign language studies*, Gallaudet University Press, v. 64, n. 1, p. 195–277, 1989. Cited on page 26.

LIEN, J. J.; KANADE, T.; COHN, J. F.; LI, C.-C. Automated facial expression recognition based on facs action units. In: IEEE. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. [S.l.], 1998. p. 390–395. Cited on page 41.

LIU, J.; LIU, B.; ZHANG, S.; YANG, F.; YANG, P.; METAXAS, D. N.; NEIDLE, C. Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. *Image and Vision Computing*, Elsevier, v. 32, n. 10, p. 671–681, 2014. Cited on page 51.

LIU, M.; LI, S.; SHAN, S.; CHEN, X. Au-aware deep networks for facial expression recognition. In: IEEE. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. [S.l.], 2013. p. 1–6. Cited on page 47.

LIU, Z.; SONG, G.; CAI, J.; CHAM, T.-J.; ZHANG, J. Conditional adversarial synthesis of 3d facial action units. *arXiv preprint arXiv:1802.07421*, 2018. Cited 4 times on pages 40, 41, 42, and 48.

LUCEY, P.; COHN, J. F.; KANADE, T.; SARAGIH, J.; AMBADAR, Z.; MATTHEWS, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. [S.l.], 2010. p. 94–101. Cited on page 45.

LUNDQVIST, D.; FLYKT, A.; ÖHMAN, A. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, v. 91, p. 630, 1998. Cited on page 44.

LYONS, M.; AKAMATSU, S.; KAMACHI, M.; GYOBA, J. Coding facial expressions with gabor wavelets. In: IEEE. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. [S.l.], 1998. p. 200–205. Cited on page 40.

LYONS, M. J.; AKAMATSU, S.; KAMACHI, M.; GYOBA, J.; BUDYNEK, J. The japanese female facial expression (jaffe) database. In: *Proceedings of third international conference on automatic face and gesture recognition*. [S.l.: s.n.], 1998. p. 14–16. Cited on page 44.

MA, C.; CHEN, L.; YONG, J. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *Neurocomputing*, Elsevier, v. 355, p. 35–47, 2019. Cited 2 times on pages 48 and 101.

MALLICK, S. Histogram of oriented gradients. *Learn OpenCV*, v. 6, 2016. Cited on page 40.

MARTINEZ, A. M. The ar face database. *CVC Technical Report24*, 1998. Cited on page 44.

MARTINEZ, B.; VALSTAR, M. F. Advances, challenges, and opportunities in automatic facial expression recognition. In: *Advances in face detection and facial image analysis*. [S.l.]: Springer, 2016. p. 63–100. Cited on page 21.

MARTINO, J. M. D.; COSTA, P. D. P.; BENETTI, A.; ROSA, L. A.; KUMADA, K. M.; SILVA, I. Building a brazilian portuguese-brazilian sign language parallel corpus using motion capture data. In: *Proceedings of Workshop of Corpora and Tools for Processing Corpora*. [S.l.: s.n.], 2016. p. 56–63. Cited on page 55.

MARTINO, J. M. D.; SILVA, I. R.; BOLOGNINI, C. Z.; COSTA, P. D. P.; KUMADA, K. M. O.; CORADINE, L. C.; BRITO, P. H. da S.; AMARAL, W. M. do; BENETTI, Â. B.; POETA, E. T. *et al.* Signing avatars: making education more inclusive. *Universal Access in the Information Society*, Springer, v. 16, n. 3, p. 793–808, 2017. Cited 4 times on pages 55, 65, 66, and 75.

MAVADATI, M.; SANGER, P.; MAHOOR, M. H. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2016. p. 1–8. Cited on page 45.

MAVADATI, S. M.; MAHOOR, M. H.; BARTLETT, K.; TRINH, P.; COHN, J. F. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, IEEE, v. 4, n. 2, p. 151–160, 2013. Cited on page 45.

MCCLEARY, L.; VIOTTI, E. Transcrição de dados de uma língua sinalizada: um estudo piloto da transcrição de narrativas na língua de sinais brasileira (lsb). *Bilinguismo dos surdos. Questões linguísticas e educacionais. Goiânia: Cânone Editorial*, p. 73–96, 2007. Cited 3 times on pages 14, 73, and 75.

MCCLEARY, L.; VIOTTI, E.; LEITE, T. de A. Descrição das línguas sinalizadas: a questão da transcrição dos dados. *ALFA: Revista de Linguística*, v. 54, n. 1, 2010. Cited 3 times on pages 14, 73, and 110.

MEHRABIAN, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, Springer, v. 14, n. 4, p. 261–292, 1996. Cited on page 41.

MEI, C.; JIANG, F.; SHEN, R.; HU, Q. Region and temporal dependency fusion for multi-label action unit detection. In: IEEE. *2018 24th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2018. p. 848–853. Cited 2 times on pages 93 and 103.

MING, K. W.; RANGANATH, S. Representations for facial expressions. In: IEEE. *Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on*. [S.l.], 2002. v. 2, p. 716–721. Cited on page 52.

MOCIALOV, B.; HASTIE, H.; TURNER, G. Transfer learning for british sign language modelling. In: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. [S.l.: s.n.], 2018. p. 101–110. Cited on page 23.

MOHR, S. The visual-gestural modality and beyond: Mouthings as a language contact phenomenon in irish sign language. *Sign Language & Linguistics*, John Benjamins, v. 15, n. 2, p. 185–211, 2012. Cited on page 51.

MOLLAHOSSEINI, A.; CHAN, D.; MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. In: IEEE. *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. [S.l.], 2016. p. 1–10. Cited 2 times on pages 90 and 93.

MURPHY-CHUTORIAN, E.; TRIVEDI, M. M. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 31, n. 4, p. 607–626, 2009. Cited on page 51.

MURTHY, V. N.; MAJI, S.; MANMATHA, R. Automatic image annotation using deep learning representations. In: ACM. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. [S.l.], 2015. p. 603–606. Cited on page 69.

NAERT, L.; REVERDY, C.; LARBOULETTE, C.; GIBET, S. Per channel automatic annotation of sign language motion capture data. In: . [S.l.: s.n.], 2018. Cited on page 23.

NEAPOLITAN, R. E. *et al. Learning bayesian networks*. [S.l.]: Pearson Prentice Hall Upper Saddle River, NJ, 2004. v. 38. Cited on page 47.

NEIDLE, C.; OPOKU, A.; DIMITRIADIS, G.; METAXAS, D. New shared & interconnected asl resources: Signstream® 3 software; dai 2 for web access to linguistically annotated video corpora; and a sign bank. *Language Resources and Evaluation*, European Language Resources Association (ELRA), 2018. Cited on page 69.

NEIDLE, C.; SCLAROFF, S.; ATHITSOS, V. Signstream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, & Computers*, Springer, v. 33, n. 3, p. 311–320, 2001. Cited on page 74.

NGUYEN, T. D.; RANGANATH, S. Towards recognition of facial expressions in sign language: Tracking facial features under occlusion. In: IEEE. *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. [S.l.], 2008. p. 3228–3231. Cited on page 51.

NGUYEN, T. D.; RANGANATH, S. Tracking facial features under occlusions and recognizing facial expressions in sign language. In: IEEE. *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. [S.l.], 2008. p. 1–7. Cited on page 52.

NISHIKIMI, R.; NAKAMURA, E.; FUKAYAMA, S.; GOTO, M.; YOSHII, K. Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism. In: IEEE. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2019. p. 161–165. Cited on page 23.

NORDQUIST, R. *Illocutionary Force in Speech Theory*. 2018. <http://www.thoughtco.com/illocutionary-force-speech-1691147>. Cited on page 72.

ONG, E.-J.; BOWDEN, R. Robust lip-tracking using rigid flocks of selected linear predictors. In: *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition*. [S.l.: s.n.], 2008. Cited on page 51.

ONG, S. C.; RANGANATH, S. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE, n. 6, p. 873–891, 2005. Cited on page 50.

PAIVA, F. A. d. S.; BARBOSA, P. A.; MARTINO, J. M. D.; WILL, A. D.; OLIVEIRA, M. R. N. d. S.; SILVA, I. R.; XAVIER, A. N. Analysis of the role of non manual expressions in intensification processes in brazilian sign language. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, SciELO Brasil, v. 34, n. 4, p. 1135–1158, 2018. Cited 7 times on pages 54, 55, 61, 64, 65, 66, and 75.

PAIVA, F. A. dos S.; MARTINO, J. M. D.; BARBOSA, P. A.; BENETTI, Â.; SILVA, I. R. Um sistema de transcrição para língua de sinais brasileira: O caso de um avatar. *Revista do GEL*, v. 13, n. 3, p. 12–48, 2016. Cited on page 75.

PANTIC, M.; ROTHKRANTZ, L. J. M. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 22, n. 12, p. 1424–1445, 2000. Cited 3 times on pages 38, 39, and 41.

PÊGO, C. F. Sinais não-manuais gramaticais da lsb nos traços morfológicos e lexicais: um estudo do morfema-boca. 2013. Cited 3 times on pages 57, 61, and 137.

PIGOU, L.; HERREWEGHE, M. V.; DAMBRE, J. Gesture and sign language recognition with temporal residual networks. In: IEEE. *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*. [S.l.], 2017. p. 3086–3093. Cited on page 51.

PIGOU, L.; OORD, A. V. D.; DIELEMAN, S.; HERREWEGHE, M. V.; DAMBRE, J. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, Springer, v. 126, n. 2-4, p. 430–439, 2018. Cited on page 51.

PIMENTA, N.; QUADROS, R. Curso de libras 1–iniciante. rev. e atualizada. *Porto Alegre: Editora*, 2008. Cited 3 times on pages 59, 60, and 63.

PIMENTA, N.; QUADROS, R. M. de. *Curso de LIBRAS 1*. [S.l.: s.n.], 2006. Cited 3 times on pages 61, 64, and 66.

PIZZIO, A. L.; QUADROS, R.; CAMPELLO, A.; REZENDE, P. Língua brasileira de sinais iii. *Apostila UFSC. Licenciatura em Letras-Libras na Modalidade a Distância, Santa Catarina*, 2009. Cited on page 59.

PRAMERDORFER, C.; KAMPEL, M. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*, 2016. Cited 3 times on pages 40, 41, and 93.

QUADROS, R. M. de; KARNOPP, L. B. *Língua de sinais brasileira: estudos lingüísticos*. [S.l.]: Artmed Editora, 2009. Cited 8 times on pages 22, 56, 60, 61, 62, 64, 65, and 137.

REZENDE, T. M. Aplicação de técnicas de inteligência computacional para análise da expressão facial em reconhecimento de sinais de libras. Universidade Federal de Minas Gerais, 2016. Cited on page 57.

REZENDE, T. M.; CASTRO, C. L. de; ALMEIDA, S. G. M. An approach for brazilian sign language (bsl) recognition based on facial expression and k-nn classifier. São José dos Campos, SP, Brazil, october 2016. Cited 7 times on pages 24, 52, 54, 57, 80, 81, and 137.

ROZADO, D.; RODRIGUEZ, F. B.; VARONA, P. Gaze gesture recognition with hierarchical temporal memory networks. In: SPRINGER. *International Work-Conference on Artificial Neural Networks*. [S.l.], 2011. p. 1–8. Cited on page 51.

ROZADO, D.; RODRIGUEZ, F. B.; VARONA, P. Low cost remote gaze gesture recognition in real time. *Applied Soft Computing*, Elsevier, v. 12, n. 8, p. 2072–2084, 2012. Cited on page 51.

RYBACH, D.; NEY, I. H.; BORCHERS, J.; DESELAERS, D.-I. T. Appearance-based features for automatic continuous sign language recognition. *Diplomarbeit im Fach Informatik Rheinisch-Westfälische Technische Hochschule Aachen*, 2006. Cited on page 69.

SAHA, S. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. [S.l.]: Towards data science, 2018. <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>. Accessed: 2020. Cited 2 times on pages 10 and 33.

SAINDON, R. J.; ESTRIN, L. H.; BRAND, D. A.; BRAND, S. *Systems and methods for automated audio transcription, translation, and transfer with text display software for manipulating the text*. [S.l.]: Google Patents, 2004. US Patent 6,820,055. Cited on page 69.

SANKARAN, N.; MOHAN, D. D.; LAKSHMINARAYANA, N. N.; SETLUR, S.; GOVINDARAJU, V. Domain adaptive representation learning for facial action unit recognition. *Pattern Recognition*, Elsevier, v. 102, p. 107127, 2020. Cited on page 103.

SANTOS, T. S. dos; XAVIER, A. N. Recursos manuais e não-manuais na expressão de intensidade em libras. *Leitura*, v. 2, n. 63, p. 120–137, 2019. Cited 5 times on pages 58, 59, 60, 63, and 137.

SARIYANIDI, E.; GUNES, H.; CAVALLARO, A. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern*

analysis and machine intelligence, IEEE, v. 37, n. 6, p. 1113–1133, 2015. Cited 5 times on pages 11, 38, 39, 40, and 41.

SCHMIDT, C.; KOLLER, O.; NEY, H.; HOYOUX, T.; PIATER, J. Using viseme recognition to improve a sign language translation system. In: *International Workshop on Spoken Language Translation*. [S.l.: s.n.], 2013. p. 197–203. Cited on page 51.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 815–823. Cited on page 49.

SEIDE, F.; LI, G.; CHEN, X.; YU, D. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: IEEE. *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. [S.l.], 2011. p. 24–29. Cited on page 69.

SHAN, C.; GONG, S.; MCOWAN, P. W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, Elsevier, v. 27, n. 6, p. 803–816, 2009. Cited on page 40.

SHAO, Z.; LIU, Z.; CAI, J.; WU, Y.; MA, L. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing*, IEEE, 2019. Cited 2 times on pages 48 and 103.

SHAO, Z.; LIU, Z.; CAI, J.; MA, L. Jaa-net: Joint facial action unit detection and face alignment via adaptive attention. *arXiv preprint arXiv:2003.08834*, 2020. Cited on page 104.

SHAO, Z.; ZOU, L.; CAI, J.; WU, Y.; MA, L. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv preprint arXiv:2001.01168*, 2020. Cited on page 103.

SHARIFARA, A.; RAHIM, M. S. M.; ANISI, Y. A general review of human face detection including a study of neural networks and haar feature-based cascade classifier in face detection. In: IEEE. *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*. [S.l.], 2014. p. 73–78. Cited 2 times on pages 10 and 39.

SHI, B.; LIVESCU, K. Multitask training with unlabeled data for end-to-end sign language fingerspelling recognition. In: IEEE. *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. [S.l.], 2017. p. 389–396. Cited on page 69.

SILVA, E. P. *Memórias autoassociativas de projeção em subespaço baseadas em estimadores robustos*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 2016. Cited 2 times on pages 30 and 31.

SILVA, E. P.; COSTA, P. D. P. Recognition of non-manual expressions in brazilian sign language. In: IEEE. *12th IEEE International Conference on Automatic Face and Gesture Recognition. Doctoral Consortium*. [S.l.], 2017. Cited on page 82.

SILVA, E. P.; COSTA, P. D. P.; KUMADA, K. M. O.; MARTINO, J. M. D.; FLORENTINO, G. A. Recognition of affective and grammatical facial expressions: a study for brazilian sign language. In: . [S.l.: s.n.], 2020. Cited 3 times on pages 81, 97, and 104.

SILVA, E. P.; COSTA, P. D. P.; KUMADA, K. M. O.; MARTINO, J. M. D. Silfa: Sign language facial action database for the development of assistive technologies for the deaf. In: . [S.l.]: IEEE, 2020. p. 382–386. Cited 9 times on pages 11, 12, 13, 59, 86, 87, 104, 105, and 106.

SILVA, E. P. d.; KUMADA, K. M. O.; COSTA, P. D. P. Analysis of facial expressions in brazilian sign language (libras). In: . [S.l.]: European Scientific Journal (ESJ), 2020. Cited on page 24.

SILVA, E. P. da; COSTA, P. D. P. Qlibras: A novel database for grammatical facial expressions in brazilian sign language. In: *X Encontro de Alunos e Docentes do DCA/FEEC/UNICAMP (EADCA)*. [S.l.: s.n.], 2017. Cited on page 81.

SIMARD, P. Y.; STEINKRAUS, D.; PLATT, J. C. Best practices for convolutional neural networks applied to visual document analysis. In: IEEE. *null*. [S.l.], 2003. p. 958. Cited on page 98.

SIMON, T.; NGUYEN, M. H.; TORRE, F. D. L.; COHN, J. F. Action unit detection with segment-based svms. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. [S.l.], 2010. p. 2737–2744. Cited on page 42.

SIMPLICIO, C.; PRADO, J.; DIAS, J. Comparing bayesian networks to classify facial expressions. In: *Proceedings of RA-IASTED, The 15th IASTED International Conference on Robotics and Applications, Cambridge, Massachusetts, USA*. [S.l.: s.n.], 2010. Cited 3 times on pages 41, 44, and 47.

SINGH, M.; MAJUMDER, A.; BEHERA, L. Facial expressions recognition system using bayesian inference. In: IEEE. *Neural Networks (IJCNN), 2014 International Joint Conference on*. [S.l.], 2014. p. 1502–1509. Cited on page 47.

SKLIAR, C. A surdez: um olhar sobre as diferenças. *Porto Alegre: Mediação*, v. 3, 1998. Cited on page 84.

SKOKI, A.; LJUBIC, S.; LERGA, J.; ŠTAJDUHAR, I. Automatic music transcription for traditional woodwind instruments sopele. *Pattern Recognition Letters*, Elsevier, v. 128, p. 340–347, 2019. Cited on page 23.

SONG, H. O.; XIANG, Y.; JEGELKA, S.; SAVARESE, S. Deep metric learning via lifted structured feature embedding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 4004–4012. Cited on page 49.

SOUZA, S. X. D. Reflexões comparativas sobre procedimentos tradutórios ao português de poemas em língua brasileira de sinais. *Mutatis Mutandis: Revista Latinoamericana de Traducción*, Universidad de Antioquia, v. 7, n. 1, p. 168–190, 2014. Cited 3 times on pages 54, 58, and 59.

SOUZA, S. X. d. *et al.* Performances de tradução para a língua brasileira de sinais observadas no curso de letras-libras. Florianópolis, SC, 2010. Cited on page 54.

SOUZA, T. A. F. d. A relação sintatico-semântica dos verbos e seus argumentos na língua brasileira de sinais (libras). Universidade Federal do Rio de Janeiro, 1998. Cited on page 71.

STOKOE, W. C. Studies in linguistics: Occasional papers 8. *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf*, 1960. Cited 2 times on pages 22 and 58.

STOKOE, W. C. Sign language structure. *Annual Review of Anthropology*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 9, n. 1, p. 365–390, 1980. Cited on page 22.

SUTSKEVER, I. *Training recurrent neural networks*. [S.l.]: University of Toronto Toronto, Ontario, Canada, 2013. Cited on page 37.

SUTTON, V. *Lessons in sign writing*. [S.l.]: SignWriting, 1995. Cited on page 70.

TAKAYAMA, N.; TAKAHASHI, H. Sign words annotation assistance using japanese sign language words recognition. In: IEEE. *2018 International Conference on Cyberworlds (CW)*. [S.l.], 2018. p. 221–228. Cited on page 23.

TANG, C.; ZHENG, W.; YAN, J.; LI, Q.; LI, Y.; ZHANG, T.; CUI, Z. View-independent facial action unit detection. In: IEEE. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. [S.l.], 2017. p. 878–882. Cited on page 49.

TAS. *Assistive Technologies for the Deaf (TAS)*. [S.l.]: University of Campinas, 2012. <http://www.tas.fee.unicamp.br/>. Accessed: 2020. Cited 2 times on pages 65 and 82.

TONG, Y.; CHEN, J.; JI, Q. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 32, n. 2, p. 258–273, 2010. Cited 2 times on pages 44 and 47.

TONG, Y.; LIAO, W.; JI, Q. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 29, n. 10, 2007. Cited on page 42.

TRAN, D.; WANG, H.; TORRESANI, L.; RAY, J.; LECUN, Y.; PALURI, M. A closer look at spatiotemporal convolutions for action recognition. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 6450–6459. Cited on page 49.

UDDIN, M. T. An ada-random forests based grammatical facial expressions recognition approach. In: IEEE. *Informatics, Electronics & Vision (ICIEV), 2015 International Conference on*. [S.l.], 2015. p. 1–6. Cited 2 times on pages 24 and 52.

VALSTAR, M. F.; JIANG, B.; MEHU, M.; PANTIC, M.; SCHERER, K. The first facial expression recognition and analysis challenge. In: IEEE. *Face and Gesture 2011*. [S.l.], 2011. p. 921–926. Cited on page 48.

VALSTAR, M. F.; SÁNCHEZ-LOZANO, E.; COHN, J. F.; JENI, L. A.; GIRARD, J. M.; ZHANG, Z.; YIN, L.; PANTIC, M. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In: IEEE. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. [S.l.], 2017. p. 839–847. Cited on page 48.

VIERA, A. J.; GARRETT, J. M. *et al.* Understanding interobserver agreement: the kappa statistic. *Fam med*, v. 37, n. 5, p. 360–363, 2005. Cited on page 107.

VIOLA, P.; JONES, M. J. Robust real-time face detection. *International journal of computer vision*, Springer, v. 57, n. 2, p. 137–154, 2004. Cited 2 times on pages 39 and 89.

VOGLER, C.; GOLDENSTEIN, S. Analysis of facial expressions in american sign language. In: *Proc, of the 3rd Int. Conf. on Universal Access in Human-Computer Interaction, Springer*. [S.l.: s.n.], 2005. Cited on page 52.

VOGLER, C.; GOLDENSTEIN, S. Facial movement analysis in asl. *Universal Access in the Information Society*, Springer, v. 6, n. 4, p. 363–374, 2008. Cited on page 52.

VOGLER, C.; LI, Z.; KANAUJIA, A.; GOLDENSTEIN, S.; METAXAS, D. The best of both worlds: Combining 3d deformable models with active shape models. In: IEEE. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. [S.l.], 2007. p. 1–7. Cited on page 52.

VOS, C. D.; KOOIJ, E. V. D.; CRASBORN, O. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands. *Language and speech*, Sage Publications Sage UK: London, England, v. 52, n. 2-3, p. 315–339, 2009. Cited on page 21.

WALAWALKAR, D. Grammatical facial expression recognition using customized deep neural network architecture. *arXiv preprint arXiv:1711.06303*, 2017. Cited 2 times on pages 24 and 52.

WALECKI, R.; RUDOVIC, O.; PAVLOVIC, V.; SCHULLER, B.; PANTIC, M. Deep structured learning for facial action unit intensity estimation. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. [S.l.], 2017. p. 5709–5718. Cited 3 times on pages 42, 47, and 48.

WANG, C.; WANG, S. Personalized multiple facial action unit recognition through generative adversarial recognition network. In: *Proceedings of the 26th ACM international conference on Multimedia*. [S.l.: s.n.], 2018. p. 302–310. Cited on page 48.

WANG, D.; BROWN, G. J. *Computational auditory scene analysis: Principles, algorithms, and applications*. [S.l.]: Wiley-IEEE press, 2006. Cited on page 69.

WANG, S.; HAO, L.; JI, Q. Facial action unit recognition and intensity estimation enhanced through label dependencies. *IEEE Transactions on Image Processing*, IEEE, 2018. Cited 2 times on pages 42 and 47.

WOODLAND, P. C.; POVEY, D. Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, Elsevier, v. 16, n. 1, p. 25–47, 2002. Cited on page 69.

WOODS, D.; FASSNACHT, C. Transana v2. 20. *Computer software] http://transana.org.. Madison, WI: The Board of Regents of the University of Wisconsin System*, 2007. Cited on page 75.

WU, Y.; JI, Q. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 3400–3408. Cited on page 47.

WU, Z.; WANG, X.; JIANG, Y.-G.; YE, H.; XUE, X. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: *Proceedings of the 23rd ACM international conference on Multimedia*. [S.l.: s.n.], 2015. p. 461–470. Cited on page 48.

XAVIER, A. N. *Descrição fonético-fonológica dos sinais da língua de sinais brasileira (LIBRAS)*. Tese (Doutorado) — Universidade de São Paulo, 2006. Cited on page 57.

XAVIER, A. N. Doubling of the number of hands as a resource for the expression of meaning intensification in brazilian sign language (libras). *Journal of Speech Sciences*, v. 1, p. 169–181, 2013. Cited on page 56.

XAVIER, A. N. *Uma ou duas? Eis a questão!: Um estudo do parâmetro número de mãos na produção de sinais da Língua Brasileira de Sinais (Libras)*. Tese (Doutorado), 2014. Cited on page 56.

XAVIER, A. N. A expressão de intensidade em libras. *Intercâmbio. Revista do Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem. ISSN 2237-759X*, v. 36, 2017. Cited 4 times on pages 56, 60, 61, and 137.

XAVIER, A. N. Análise preliminar de expressões não-manuais lexicais na libras. *Intercâmbio. Revista do Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem. ISSN 2237-759X*, v. 40, 2019. Cited 5 times on pages 22, 57, 60, 64, and 65.

XAVIER, A. N.; BARBOSA, P. Com quantas mãos se faz um sinal? um estudo do parâmetro número de mãos na produção de sinais da língua brasileira de sinais (libras). *Todas as Letras-Revista de Língua e Literatura*, v. 15, n. 1, 2013. Cited on page 56.

XAVIER, A. N.; BARBOSA, P. A. Diferentes pronúncias em uma língua não sonora? um estudo da variação na produção de sinais da libras. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, SciELO Brasil, v. 30, n. 2, p. 371–413, 2014. Cited 2 times on pages 56 and 63.

XIONG, W.; WU, L.; ALLEVA, F.; DROPPO, J.; HUANG, X.; STOLCKE, A. The microsoft 2017 conversational speech recognition system. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2018. p. 5934–5938. Cited on page 69.

XU, Y. Implement long short-term memory recurrent neural network on grammatical facial expression recognition. 2017. Cited 2 times on pages 24 and 52.

ZENG, Z.; PANTIC, M.; ROISMAN, G. I.; HUANG, T. S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 31, n. 1, p. 39–58, 2009. Cited 2 times on pages 38 and 39.

ZEYER, A.; DOETSCH, P.; VOIGTLAENDER, P.; SCHLÜTER, R.; NEY, H. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. [S.l.], 2017. p. 2462–2466. Cited on page 69.

ZHANG, M.-L.; ZHOU, Z.-H. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, IEEE, v. 26, n. 8, p. 1819–1837, 2014. Cited on page 47.

ZHANG, X.; YIN, L.; COHN, J. F.; CANAVAN, S.; REALE, M.; HOROWITZ, A.; LIU, P.; GIRARD, J. M. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, Elsevier, v. 32, n. 10, p. 692–706, 2014. Cited on page 45.

ZHANG, Y.; ZHAO, R.; DONG, W.; HU, B.-G.; JI, Q. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 7034–7043. Cited on page 42.

ZHANG, Z.; GU, J. Facial affect recognition in the wild using multi-task learning convolutional network. *arXiv preprint arXiv:2002.00606*, 2020. Cited on page 49.

ZHAO, G.; HUANG, X.; TAINI, M.; LI, S. Z.; PIETIKÄINEN, M. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, Elsevier, v. 29, n. 9, p. 607–619, 2011. Cited on page 44.

ZHAO, K.; CHU, W.-S.; MARTINEZ, A. M. Learning facial action units from web images with scalable weakly supervised clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 2090–2099. Cited 5 times on pages 42, 46, 47, 49, and 97.

ZHAO, K.; CHU, W.-S.; TORRE, F. De la; COHN, J. F.; ZHANG, H. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, IEEE, v. 25, n. 8, p. 3931–3946, 2016. Cited on page 48.

ZHAO, K.; CHU, W.-S.; ZHANG, H. Deep region and multi-label learning for facial action unit detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 3391–3399. Cited 5 times on pages 40, 41, 42, 46, and 47.

ZHOU, Y.; PI, J.; SHI, B. E. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In: IEEE. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. [S.l.], 2017. p. 872–877. Cited on page 49.

ZWILLINGER, D. *CRC standard mathematical tables and formulae*. [S.l.]: CRC press, 2002. Cited on page 34.

APPENDIX A – Protocol for Bibliographic Survey of Libras' Facial Expressions

The survey on Libras facial expressions was carried out from the identification and evaluation of works[1] that included analysis of facial expressions and their functions within this linguistic system.

While the survey did not implement a typical systematic review approach, with independent reviewers, we established the following guidelines that oriented the review:

A.1 Guidelines

In the idealization process, we have defined a set of rules for searching facial expressions of Brazilian sign language and we established them below.

Objective: Find and evaluate works that bring linguistic analysis of facial expressions and their functions within the Brazilian sign language.

Research questions: To build a classification model, some information is needed for the labeling of images. So, below are some questions that we needed to answer.

1. How many, and what are the facial expressions found in Libras?
2. What are the functions of different facial expressions in Libras?
3. How facial expressions are described in Libras?

Source selection: In order to be scientifically sounded, the works must be on scientific bases. Also, available online. So we choose two search sources:

1. Google Scholar (https://scholar.google.com)
2. Scientific Eletronic Library Online (SciELO) (https://scielo.org/)

Search descriptor: The following English and Portuguese search sentences were derived from articles indicated by researchers of linguistic field.

Search string in English:

- Non-manual markers AND Libras OR Brazilian Sign Language

[1] We consider works as articles, dissertations, thesis and, conference papers.

- Facial Expression AND Libras OR Brazilian Sign Language
- Grammatical facial expression AND Libras OR Brazilian Sign Language
- Grammatical facial expression OR non-manual expression OR nonmanuals

Search string in Portuguese:

- Sinais Não-Manuais AND Libras OR Língua Brasileira de Sinais
- Expressões Faciais AND Libras OR Língua Brasileira de Sinais
- Expressão facial gramatical AND Libras OR Língua Brasileira de Sinais
- Expressão facial gramatical OR Expressão não-manual

Inclusion criteria: To be included in the full reading, works must contain:

- Description of facial expression in Libras.
- Characterization of facial expression in Libras.
- Description of functions of facial expression in Libras.
- Detailed described approach for facial expressions mentioned.
- Considers at least one facial movement.

Exclusion criteria: If any statements below are true, the works was excluded from our survey.

- No mention of the method of studying the expressions.
- Incorporate only body movements and not facial expressions or head movement.
- Does not mention Libras.
- Articles from the same author with the same theme and older with fewer differences between them.

Data selection and extraction: The descriptors were included in the chosen search engines. The primary reviewer read the title, the abstract, the sections, and the conclusion. After applying the inclusion and exclusion criteria, the included works were designated to a full reading. The reviewer made a summarizing of the principal results, separated the facial expression described, the type of analysis made, type of methodology, and if any kind of corpus was created or used.

A.2 Development

In the last search for the first source search, the results were ordered by relevance for each string. If there were more then one page of results, we considered until that the six consecutive turns out of results were not relevant for our research. In the second source search, there was only one page of results for each string. The pre-selection process consisted of reading the title, the abstract, the structure of sections, and the conclusion of each work. Later the inclusion and exclusion criteria were applied. A number of works were selected for full reading analysis, from both Scielo and Google Scholar, some coincident. In fact, were analyzed the following works: Arrotéia (2005), Freitas (2011), Freitas *et al.* (2014), Araujo (2013), Felipe (2013), Pêgo (2013), Almeida (2014), Rezende *et al.* (2016), Xavier (2017), Santos e Xavier (2019). Additionally, to the works found, books and other references indicated by professionals in the field were included in the analysis as follows: Capovilla *et al.* (2017), Quadros e Karnopp (2009). In Chapter 3 was made a description summarizing some results, separated the facial expression described, the type of analysis made, type of methodology, and if any corpus was created or used.

APPENDIX B – Association between Facial Action Coding System and Facial Expressions of Brazilian Sign Language.

In this attached chapter we present tables with images illustrating associations between Facial Action Coding System and facial expressions of Brazilian Sign Language. In Table B.1, Table B.2 and, Table B.3 we present the simple expressions. While in Table B.4 we combine the simple expressions forming compound expressions.

Table B.1 – Association between Libras facial expressions and FACS with images comprehending movements of the head

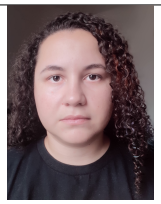
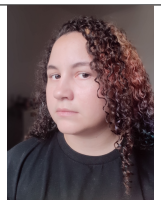
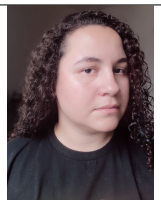
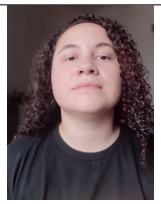
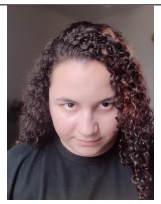
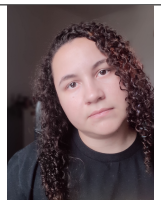
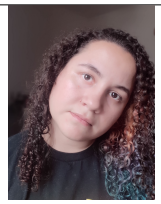
Head					
Libras taxonomy	Neutral	Balancing sideways (no)		Balance back and forth(yes)	
AU	0	51	52	53	54
Description	No face action.	Head left	Head right	Head up	Head down
Image					
Libras taxonomy	Tilt back	Quick nod	Tilt to the side		Forward lean
AU	53	54	55	56	57
Description	Head up	Head down	Head Tilt left	Head Tilt right	Head foward down
Image					

Table B.2 – Association between Libras facial expressions and FACS with images comprehending movements of the upper part of the face

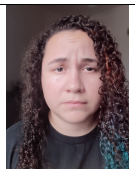
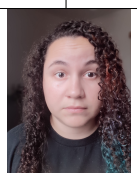
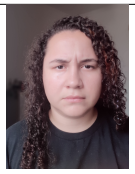
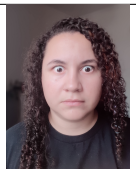
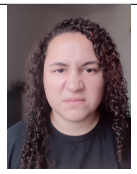
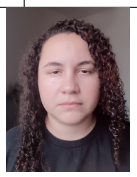
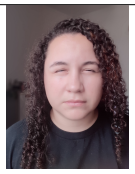
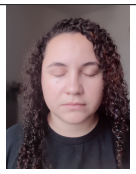
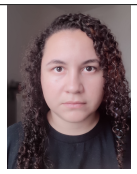
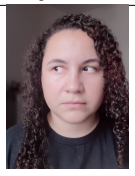
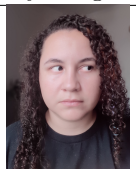
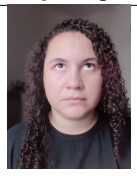
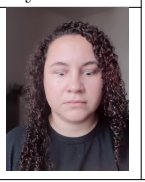
Upper face					
Libras taxonomy	Joined eyebrows	Raised eyebrows		Left / Right eyebrow raised	
AU	1	1	2	2L / 2R	
Description	Inner Brow Raiser	Inner Brow Raiser	Outer Brow Raiser	Left / Right Outer Brow Raiser	
Image					
Libras taxonomy	Frown	Wide open eyes	Nose wrinkle	Slightly closed eyes	
AU	4	5	9	41	42
Description	Brow Lowerer	Upper Lid Raiser	Nose Wrinkler	Lid droop	Slit
Image					
Libras taxonomy	Eyes tight	Closed eyes	Left / Right eye closed		Look at the speaker
AU	44	45	46		-
Description	Squint	Eyes Closed	Wink		-
Image					
Libras taxonomy	Direct the eyes				
AU	61	62	63	64	
Description	Eyes left	Eyes Right	Eyes Up	Eyes Down	
Image					

Table B.3 – Association between Libras facial expressions and FACS with images comprehending movements of the lower part of the face

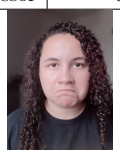
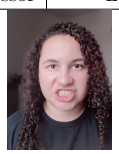
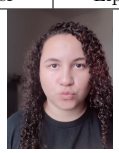
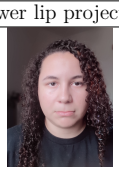
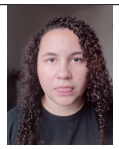
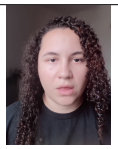
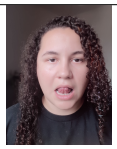
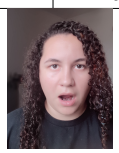
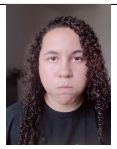
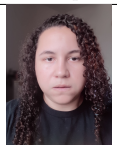
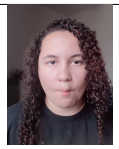
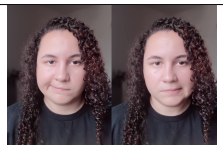
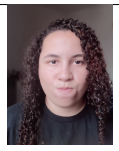
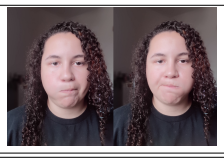
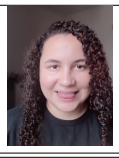
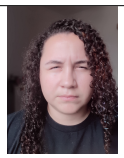
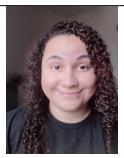
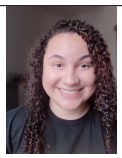
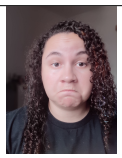
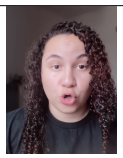
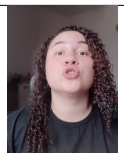
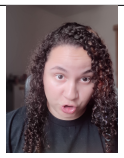
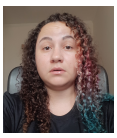
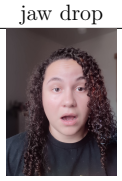
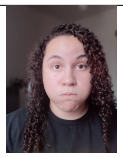
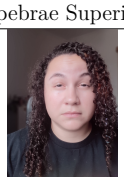
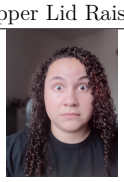
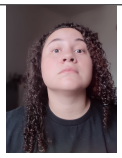
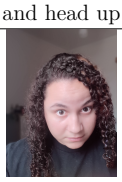
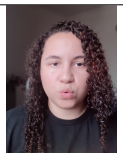
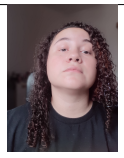
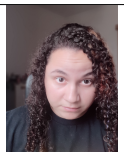
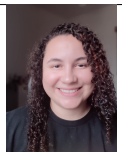
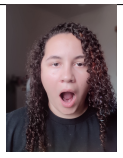
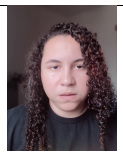
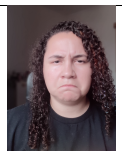
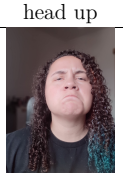
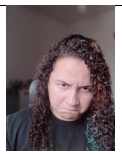
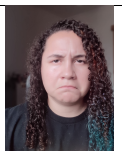
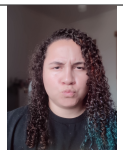
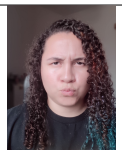
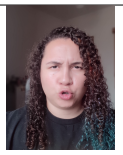
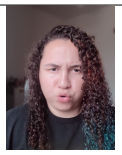
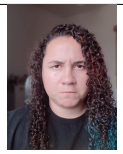
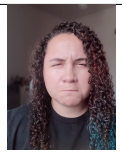
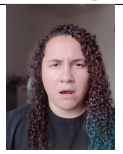
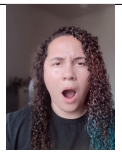
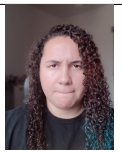
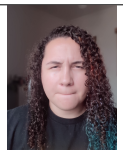
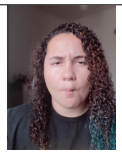
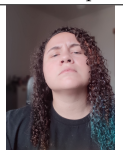
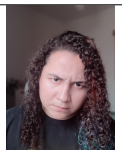
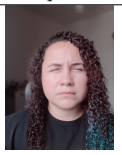
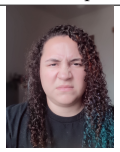
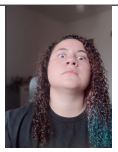
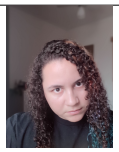
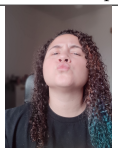
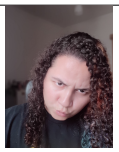
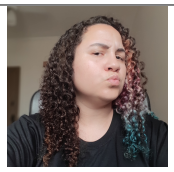
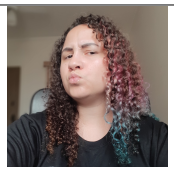
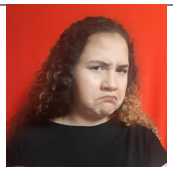
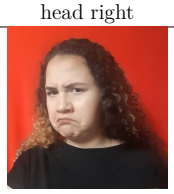
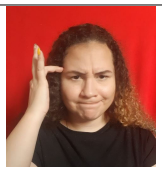
Lower face					
Libras taxonomy	Crooked mouth up	Crooked mouth down		Clenched teeth	
AU	12	15	17	16	22
Description	Lip corner puller	Lip corner depressor	Chin raiser	Lower Lip Depressor	Lip Funneler
Image					
Libras taxonomy	Projected lips		Lips pressed (closed mouth)	Contracted lips	Contraction of the upper lip
AU	18	23	24	28	17 + 28
Description	Lip Puckerer	Lip Tightener	Lip Pressor	Lip suck	Lower lip projected
Image					
Libras taxonomy	Tongue in lisp position	Swinging alveolar tongue	Sibilant tongue	Tip of the tongue touching the lips	Contraction of the lower lip
AU	19	86	87	88	
Description	Tongue show				Lip bite
Image					
Libras taxonomy	Open mouth		Inflated cheeks	Cheek blowing	Contracted cheeks
AU	25	26	33	34	35
Description	Lips part	Jaw Drop	Cheek blow	Cheek puff	Cheek suck
Image					
Libras taxonomy	Crooked mouth up laterally		Crooked mouth down laterally		Run the tongue against the lower part of the cheek
AU	AU12R / AU12L		AU15R / AU15L		19 + 20
Description	-		-		
Image					
Libras taxonomy	Only left / right cheek inflated		Chewing motion		Bite on the tongue
AU	AU34R / AU34L		81		16 + 19 + 22
Description	-		-		-
Image					

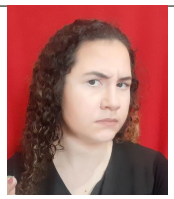
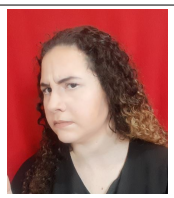
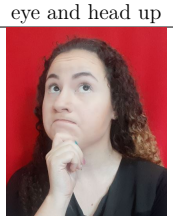
Table B.4 – Association between Libras facial expressions and FACS with images comprehending composed movements of the face

Compound facial expressions					
Libras taxonomy	Frown, Slightly closed eyes	Raised eyebrows, crooked mouth up	Raised eyebrows and smile with apparent teeths	Raised eyebrows, crooked mouth down	Raised eyebrows, Projected lips and open mouth
AU	4 + 41 + 42	1 + 2 + 12	1 + 2 + 12 + 25	1 + 2 + 15 + 17	1 + 2 + 22 + 25
Description	Brow Lowerer and Eyes Closed	Inner and Outer Brow Raiser, Lip corner puller	Inner and Outer Brow Raiser, crooked mouth up, open mouth	Inner and Outer Brow Raiser, Lip corner depressor and chin raiser	Inner and Outer Brow Raiser, Lip Puckerer and Tightener
Image					
Libras taxonomy	Raised eyebrows, projected lips, open mouth, and head balance back and forth		Raised eyebrows, and open mouth		Raised eyebrows and Inflated cheeks
AU	1+2+22+25+53	1+2+22+25+54	1 + 2 + 25	1 + 2 + 25 + 26	1 + 2 + 33
Description	-	-	Inner and Outer Brow Raiser, Lips apart	Inner and Outer Brow Raiser, jaw drop	-
Image					
Libras taxonomy	Raised eyebrows, and slightly closed eyes	Raised eyebrows, and wide open eyes	Raised eyebrows, wide open eyes and head balance back and forth		Projected lips, open mouth and inflated cheeks
AU	1+2+41+42	1+2+5	1+2+5+53	1+2+5+54	22+25+33
Description	Inner and Outer Brow Raiser, Relaxation of Levator Palpebrae Superioris	Inner and Outer Brow Raiser and Upper Lid Raiser	Inner and Outer Brow Raiser Upper Lid Raiser, and head down	Inner and Outer Brow Raiser Upper Lid Raiser, and head up	-
Image					
Libras taxonomy	Raised eyebrows, and head balance back and forth		Crooked mouth up, and open mouth		Open mouth
AU	1+2+53	1+2+54	12+25	12+25+26	25+27
Description	Inner and Outer Raiser, head up	Inner and Outer Raiser, head down	Smile with aparent teeth	Smile with aparent teeth and jaw drop	Masetter; Temporal and Internal Pterygoid relaxed
Image					

Libras taxonomy	Open mouth and inflated cheeks	Frown and crooked mouth down	Frown, crooked mouth down and head Balance back and forth.		Frown and crooked mouth down
AU	25+33	4+15+17	4+15+17+53	4+15+17+54	4+17
Description	-	Brow Lowerer, Lip corner depressor and Chin Raiser	Brow Lowerer, Lip corner depressor, Chin Raiser and head up	Brow Lowerer, Lip corner depressor, Chin Raiser and head down	Brow Lowerer and Chin Raiser
Image					
Libras taxonomy	Frown and projected lips	Frown, projected lips and inflated cheeks	Frown, projected lips and open mouth	Frown, projected lips, open mouth and inflated cheeks	Frown, open mouth, and inflated cheeks
AU	4+18+22	4+18+33	4+22+25	4+22+25+33	4+25+33
Description	Brow Lowerer, Incisivii labii superioris and Incisivii labii inferioris	Brow Lowerer, Lip pucker, and Cheeks blowing	Brow Lowerer and Orbicularis oris	Brow Lowerer and Cheeks blowing	-
Image					
Libras taxonomy	Frown and lips pressed	Frown, slightly closed eyes, and lips pressed	Frown and open mouth	Frown and open mouth	Frown and contracted lips
AU	4+24	4+24+41+42	4+25+26	4+25+27	4+28
Description	Brow Lowerer and closed mouth		Brow Lowerer and Jaw drop	Brow Lowerer and Mouth Stretch	
Image					
Libras taxonomy	Frown, contracted lips and slightly closed eyes	Frown and contracted cheeks	Frown and head balance back and forth		Frown, slightly closed eyes and eyes tight
AU	4+28+41	4+35	4+53	4+54	4+41+42+44
Description		Brow Lowerer and cheek suck	Brow Lowerer and head up	Brow Lowerer and head down	Brow Lowerer and Squint
Image					

Libras taxonomy	Frown and nose wrinkle	Wide open eyes, and head balance back and forth		Frown, projected lips, and head balance back and forth	
AU	4+9	5+53	5+54	4+18+22+53	4+18+22+54
Description	Brow Lowerer, Levator labii superioris alaquae nasi	Upper Lid Raiser and head up	Upper Lid Raiser and head down	Brow Lowerer, Lip Puckerer and head up	Brow Lowerer, Lip Puckerer and head down
Image					

Libras taxonomy	Frown, projected lips and head balacing sideways		Frown, crocked mouth down and head balacing sideways		Frown, Crocked mouth and Contracted lips
AU	4+18+22+51	4+18+22+52	4+15+17+51	4+15+17+52	4+13+28
Description	Brow Lowerer, Lip Puckerer and Funneler, and head left	Brow Lowerer, Lip Puckerer and Funneler, and head right	Brow Lowerer, Lip corner depressor, Chin Raiser and head left	Brow Lowerer, Lip corner depressor, Chin Raiser and head right	Brow Lowerer, Levator anguli oris (Caninus), Orbicularis oris
Image					

Libras taxonomy	Frown, projected lips and head balacing sideways		Frown, crocked mouth down and head balacing sideways		Raised eyebrows, crocked mouth down, eyes and head up
AU	4+51	4+52	15+17+51	15+17+52	1+2+15+17+53+63
Description	Brow Lowerer and head left	Brow Lowerer and head right	Lip corner depressor, Chin Raiser and head left	Lip corner depressor, Chin Raiser and head right	Outer Brow Raiser, Lip corner depressor, Chin Raiser, eye and head up
Image					

Libras taxonomy	Raised eyebrows, crocked mouth down and head balacing sideways		Raised eyebrows and head balacing sideways	
AU	1+2+15+17+51	1+2+15+17+52	1+2+51	1+2+52
Description	Brow Lowerer and head left	Brow Lowerer and head right	Lip corner depressor, Chin Raiser and head left	Lip corner depressor, Chin Raiser and head right
Image